Relational Integration in Working Memory: Determinants of Effective Task Performance and Links to Individual Differences in Fluid Intelligence

Joel Edward Bateman
BPsych (Hons. I)

School of Psychology
The University of Sydney
Sydney, New South Wales
AUSTRALIA

A thesis submitted to fulfil requirements for the degree of
Doctor of Philosophy

11th May 2020

STATEMENT OF ORIGINALITY AND APPROVAL

This statement certifies that this thesis has not been submitted for a higher degree or for any other purposes to any other university of institution. To the best of my knowledge and belief, the content of this thesis is my own work.

The data obtained through the research presented in this thesis was collected with approval by the Human Research Ethics Committee at the University of Sydney (through protocols 2016/690, 2015/275 and 2014/287).

Joel Edward Bateman

PhD Candidate

AUTHORSHIP ATTRIBUTION STATEMENT

Some data from Chapters III and IV were partially presented as: [Bateman, J. E., Birney, D. P., & Loh, V. (2017). *Exploring functions of working memory related to fluid intelligence: Coordination, relational integration, and access.* Paper presented at the 39th Annual Meeting of the Cognitive Science Society, London, UK.] I designed the study, analysed the data, constructed the poster, and wrote the manuscript version of the paper.

Chapter IV of this thesis is published as: [Bateman, J. E., & Birney, D. P. (2019). The link between working memory and fluid intelligence is dependent on flexible bindings, not systematic access or passive retention. *Acta Psychologica*, *199*, *102893*]. I designed the study, contributed to the analyses of the data, and wrote the drafts of the manuscript.

Chapter VI of this thesis is published as: [Bateman, J. E., Thompson, K., A., & Birney, D. P. (2019). Validating the relation monitoring task as a measure of relational integration and predictor of fluid intelligence. *Memory & Cognition*, *47* (8), 1457-1468.]. I designed the study, analysed the data, and wrote the drafts of the manuscript.

Chapter VII of this thesis is partially published as: [Bateman, J. E., Ngiam, W. X. Q., & Birney, D. P. (2018). Relational encoding of objects in working memory: Change detection performance is better for violations in group relations. *PLoS ONE, 13 (9), e0203848*]. I designed the study, contributed to the analyses of the data, and wrote the drafts of the manuscript.

Joel E. Bateman, 12th January 2020.

As supervisor for the candidature upon which this thesis is based, I can confirm that the authorship attribution statements above are correct.

Damian P. Birney, 12th January 2020.

Table of Contents

Abstract

Working memory is a critical system of human cognition, providing a conscious stream of thought that allows us to focus attention, store and manipulate temporary information, and flexibly solve complex problems. Although traditionally seen as a multi-componential system with distinct capacity-limited stores (Baddeley & Hitch, 1974), there is a growing consensus that working memory is a more dynamic, attentional-based system limited by the ability to both maintain and disengage from memory representations. Central to this maintenance and disengagement is the integration of representations by binding them into established or novel relations – a process termed *relational integration*. Working memory tasks are often linked to higher-order abstract reasoning (fluid intelligence) tasks which requires abstraction of relations; and the capacity for relational integration is prevalent throughout comparative cognition. Despite this, the nature of relational integration within working memory is not well understood. This is at least in part due to the difficulty in quantifying unique relational integration demands, separately from well-established passive storage theories and attentional control theories, where predicted outcomes often coincide. The current project aims to understand the nature of relational integration in working memory, identifying aspects of relational integration which contribute to successful task performance on working memory and fluid intelligence tasks. To this end, several studies are conducted which investigate determinants of relational integration including complexity, salience, and systematicity. Consistent evidence emerges that indicates the ability to establish, maintain, and dissolve multiple strong and flexible bindings is the best predictor of task performance on relational integration tasks; and can predict well-established abstract reasoning tasks over-and-above classic working memory tasks which emphasize attentional control demands or at least, a demarcation of storage and processing.

RELATIONAL INTEGRATION IN WORKING MEMORY: DETERMINANTS OF EFFECTIVE TASK PERFORMANCE AND LINKS TO INDIVIDUAL DIFFERENCES IN FLUID INTELLIGENCE

I. INTRODUCTION TO RELATIONAL INTEGRATION

1.1. Humans as relational thinkers

Humans are capable of interpreting complex relationships between both similar and dissimilar, real and abstract objects and concepts. We can piece together information within a novel scenario with other information attained over a lifespan, creating a complex constellation of relationships. These relationships can be as specific as the difference in *love* between *John loves Mary* and *John loves Fido* (Hummel & Holyoak, 2001). Although other species have demonstrated remarkable abilities to extract similarities and differences, there is no convincing evidence that this abstraction is anything more than complex behavioural learning or feature matching (Penn, Holyoak, & Povinelli, 2008). Conversely, human cognition can extract subtle relations, or generate relations between objects with no intrinsic similarities. This ability often manifests more as a *tendency*, seeing patterns in unrelated stimuli, as in apophenia or confirmation bias (Waterstone, 2007). Although this occasionally causes issues (Paul, Monda, Olausson, & Reed-Daley, 2014), our relational abilities are often effective and unlock limitless potential for learning new information through analogy (Penn et al., 2008).

The tendency to think via relations has also enabled us to bypass cognitive limitations. In contrast to brief sensory memory (lasting under one second) which has a large capacity to accommodate perceptual experiences (Sperling, 1960), short-term memory (STM) capacity (in the realm of seconds and minutes) is severely limited (Cowan, 2017) to only a few pieces of information. Our ability to generate meaningful relations among otherwise independent representations allows us to circumvent this original capacity and enables more meaningful interpretation of relationships between the representations. This is typically referred to as *chunking* (Cowan, 2001; Miller, 1956). In typical measures of STM, retaining elements in a series is seen as a measure of capacity (Colom, Rebollo, Abad, & Shih, 2006) and chunking is a natural strategy to subvert this capacity (Feigenson & Halberda, 2008). Often, effort must be taken to purposely disrupt chunking by ensuring to-be-remembered elements have no intrinsic relation (Portrat, Guida, Phénix, & Lemaire, 2016) or by diverting processing by requiring some manipulation of the series (such as recall in reverse order; Richardson, 2007). This manipulation is a defining feature of *working memory*, which is often seen as a dual-module system for active processing (manipulation) and passive storage (maintenance) (Baddeley & Hitch, 1974). For instance, remembering the digits *27* and *18* would primarily involve passive storage (perhaps intermixed with more 'active' rehearsal; Tan & Ward, 2008), while summing them together would also require active processing in order to increment (a form of manipulation) the operands and generate the outcome (Dehaene, 1992). Thus, mental arithmetic is often seen as a quintessential working memory task (DeStefano & LeFevre, 2004). Although there is no single agreed definition for working memory (Cowan, 2017), there is a growing consensus that the ability to construct relations between representations – *relational integration* – is a critical determinant of working memory task performance (Oberauer et al., 2018). There is also considerable emerging evidence to suggest that the capacity for relational integration is the strongest predictor of higher-order abstract reasoning (Chuderski, 2014; Oberauer, Süß, Wilhelm, & Wittman, 2008).

This thesis argues that relational integration is the foundation of working memory. Even the most rudimentary storing of elements in STM requires relational integration: temporary binding of memory representations to a place within a relational structure (Oberauer, 2009a; Robin & Holyoak, 1995). Together, a set of integrated bindings allows us to draw associations between representations (Oberauer, Süß, Wilhelm, & Sander, 2007). The following sections of this chapter outlines the current interpretation of working memory being a system based on relational integration. It will also discuss the much-cited overlap between working memory and higher-order “fluid intelligence” tasks which involve abstract reasoning (Ackerman, Beier, & Boyle, 2005; Shipstead, Harrison, & Engle, 2016). The remainder of this chapter then describes the objectives and plan of the current project: investigating factors contributing to effective working memory performance and links to abstract reasoning. In Chapter II, the background theories and research contributing to a new definition of working memory as a system for relational integration are explored.

1.2. Working Memory as a system for Relational Integration

Section 1.1. outlined STM as a passive storage system and highlighted processing demands as a defining feature that extends STM into *working memory* (WM), a system that accounts for both maintenance and manipulation (Baddeley & Hitch, 1974) of temporarily activated representations. In addition to maintenance and manipulation (storage and processing), theories of WM must also account for the important role of *attention* (Baddeley, 1993), which distinguishes (i) centrally focused representations, said to be within ‘central attention’ (Broadbent, 1958); from (ii) unfocused representations that are held active (Cowan, 1995; Fougnie, 2008; Oberauer, 2002) which can be immediately accessed by central attention (as they are just outside the focus of attention); from (iii) inactive representations held in long-term memory which require deliberate or primed retrieval to be activated (Oberauer & Hein, 2012). Although it may be easier to think of each of the three functions of WM (storage, processing, attention) as operating distinctly (as was the norm for multi-componential models, e.g., Baddeley & Hitch, 1974), a primary argument of this thesis is that the three functions can be understood together through a theory of working memory where relational integration is its foundation. At the very least, it is argued that the storage and processing functions emerge through relational integration. For instance, recalling a list of serially ordered digits appears to be a theoretically ‘simple’ storage task, involving storing each digit as an independent representation within a temporary buffer that can be easily accessed (Unsworth & Engle, 2006). Similarly, a system based on relational integration allows us to bind each digit to its place within the order but also allows us to manipulate the order by binding the digits to new places. We propose that the capacity for relational integration is based on the ability to construct multiple strong and flexible bindings. Bindings must be *strong* in the face of interfering information (Lewandowsky, Geiger, Morrell, & Oberauer, 2010; Oberauer & Lewandowsky, 2008) that could otherwise degrade bindings and compromise relations (for instance, recalling a list of digits in serial order while ignoring irrelevant letter distractors). Bindings must also be *flexible* to accommodate shifting and updating of element relations (Kessler & Oberauer, 2014) that may be required (for instance, rearranging a randomly-ordered list of digits in arithmetically ascending order). The number of strong and flexible bindings that can be active at a single time is an indication of *binding capacity*, the relational integration equivalent to storage capacity in traditional views of working memory (Conway, Jarrold, Kane, Miyake, & Towse, 2007). Whether binding capacity varies between individuals; or whether binding capacity is fixed, and the variation occurs in the flexibility and strength of these bindings, remains to be seen.

Binding and relational integration have been used generally to refer to drawing associations between elements within memory (e.g., Olsen et al., 2015; Sluzenski, Newcombe, & Kovacs, 2006), though the current project necessitates a theoretical account of relational integration. Relational integration involves generating a relational structure by binding an element to a role within a relation (Halford, Wilson, & Philips, 1998; Hummel & Holyoak, 2001; Oberauer, 2009a). The *role* signifies the role being played by the element in the association.[1] For instance, when *John loves Mary*, John is in the *lover* role and Mary is in the *loved* role (Hummel & Holyoak, 2001). Several roles constitute a relational set (although one-dimensional or 'unary' sets consisting of only one role, such as an attribute, may also be considered relations, Halford et al., 1998). The number of roles that make up a set signifies the dimensionality of the relation (Halford et al., 1998). For instance, *loves* is a two-dimensional relation consisting of a *lover* and a *loved*. The result of relational integration is the instantiation of a relational instance that allows us to comprehend associations between multiple elements in WM.[2] That is, the otherwise independent elements (John and Mary) have been *integrated* into a meaningful relation, allowing us to comprehend that John is the one loving Mary (because he is in the *lover* role, an active initiating agent), and that Mary may or may not reciprocate this love (because she is in the *loved* role, a passive target). In this way, relational integration is termed as such not because it is about integrating relations (as in analogy) but because it involves integration of the relational kind (*relational* is an adjective, not a nominalisation). It should also be noted that, although *binding* may be used shorthand to refer to binding elements together (as in, “binding a series of digits together”), this usage belies the actual binding process, which involves binding elements to roles. Thus, a more accurate statement would be “binding digits into a series”, where the underlying process involves each digit being bound to a location in the series.

In Oberauer’s (2009a) model of working memory, which draws on connectionist architectures involving networks of interconnected nodes, elements are represented in *content* nodes and roles are represented in *context* nodes. Content nodes are bound to context nodes so that the element and its role may be represented together. Both content and context nodes are extremely versatile (Oberauer, 2009a), allowing for the free generation of virtually any

[1] A role can also be referred to as a *slot* in filler-slot terminology (Halford et al., 1998; Robin & Holyoak, 1995).

[2] For the purposes of the current project, we limit the scope of relational integration to WM, rather than e.g., episodic memory, where it has also featured (Olson & Newcombe, 2014).

association between any set of elements. For instance, we may instantiate a relation between *rat* and *mouse* by binding each of them to the taxonomical order of *rodent*, which is itself a feature of both rats and mice. More broadly, we can also relate *rat* and *mouse* by binding them to the taxonomical class of *mammals*. Alternatively, we may relate *rat* and *mouse* by binding them to an anatomical feature, like the fact they both have tails. If they are in our immediate vicinity, we can recognize that the rat is further from us in physical space, by binding the rat to a *closer* spatial context and the mouse to *further*. We could also relate the rat and mouse by the fact that they both feature in this sentence, or by the fact they are both being represented here in English orthography. There are virtually infinite ways to construct a relation, and relational integration accommodates this by allowing any element to be bound to any role (Oberauer, 2009a). The bound roles inform of the association between the elements they are bound to, such as similarities (each being bound to a *rodent* role) or differences (mouse being bound to *smaller*).

Our flexibility in binding means it is possible to construct novel relations, such as a mouse being bound to *larger* and a rat being bound to *smaller*, though experience with the contrary relation (i.e., mice actually being smaller than rats) means this binding is more effortful as it must contend with *schemas* that have been well established in LTM through repeated exposure. As a simple example of schemas, depending on the goal of the task, rearranging the letters I-B-F could involve retrieving a schema for alphabetical order (B-I-F) *or* by retrieving a schema of a familiar acronym (F-B-I). Constructing novel relations without relying on schemas is often critical for solving novel problems. For instance, recalling the randomly ordered list of letters I-B-F requires constructing then maintaining a novel relation representing the temporal order *first-second-third* (if the random order is otherwise meaningless). Although constructing novel relations is often a critical component of the task, retrieving established schemas may also be required depending on task demands. Although this appears akin to the contrast between fluid and crystallized intelligence (Horn & Cattell, 1966), there is evidence to suggest that schemas can be strategically developed over the course of reasoning tasks (Thompson, Prowse Turner, & Pennycook, 2011). Both are required for day-to-day functional WM. Although binding is broadly flexible for the reasons discussed, individual differences in the flexibility of binding may be considered in the ability to contend with highly established – but unhelpful – schemas. Figure 1.1 provides a visualisation of how a recall task can be handled using either retrieval of a relevant schema or by constructing a novel relation. Although either can work for this particular task, they each have advantages and disadvantages that are amplified according to the task format. For instance, constructing a novel relation representing temporal order is more likely to be lost to interference as it has not yet been committed to a well-established schema in LTM; though if the words lack any meaningful relation, there may be no schema that can be retrieved. Well-established schemas may also cause intrusions from related non-target words.

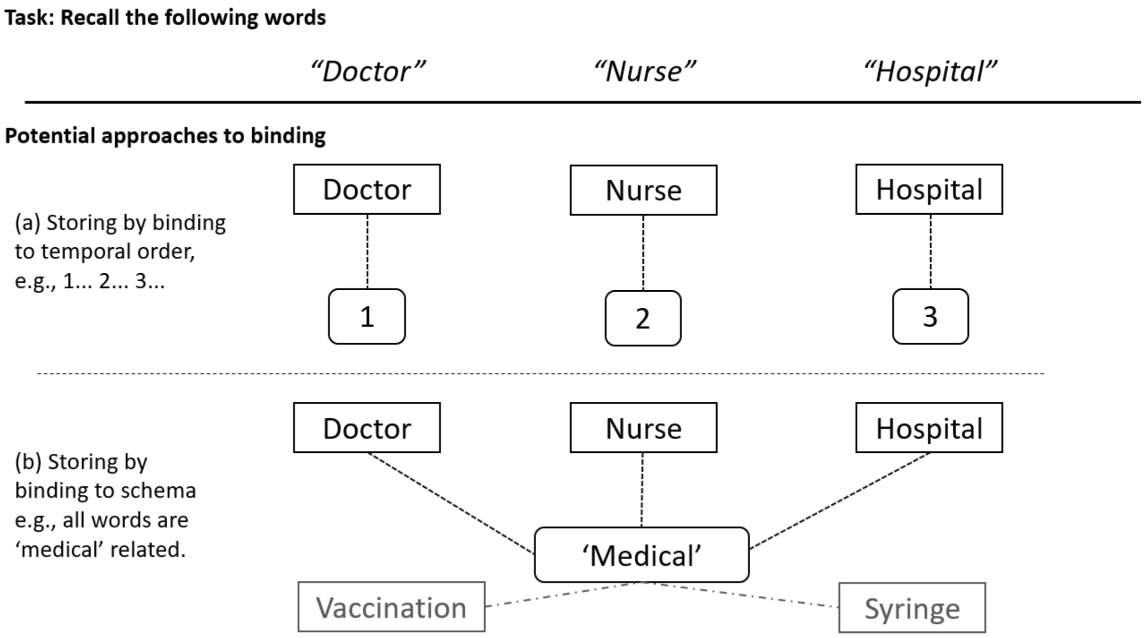

Task: Recall the following words
"Doctor"
"Nurse"
"Hospital"
Potential approaches to binding
Doctor
Nurse
Hospital
(a) Storing by binding to temporal order, e.g., 1… 2… 3…
1
2
3
Doctor
Nurse
Hospital
(b) Storing by binding to schema e.g., all words are 'medical' related.
'Medical'
Vaccination
Syringe

Figure 1.1. Demonstration of how a simple recall task elicits relational integration and can be solved by either (a) constructing a novel relation (e.g., temporal order) or (b) by retrieving a schema set from LTM that can be applied to the relation. Instantiation of temporal order among the elements involves binding each element to a context relating its temporal position (e.g., 1st, 2nd, …). In this example, retrieval of a semantic schema allows each word to be bound to a *medical* context, propagating activation of elements at recall. Maintaining only one unique context representation ('medical') reduces the cognitive load of the task, but the overreliance on this single context in lieu of more specific contexts (like temporal order) can result in intrusions by lure words such as *syringe*, which have some tangential activation.

The contrast between construction of novel relations and retrieval of established schema sets can explain past research. Consider a dual task paradigm where the primary task is to remember a list of words and the secondary task involves verifying the grammatical veracity of sentences. From a relational integration perspective of WM, we would predict that similarities between to-be-remembered words in the primary task and distracting words in the secondary task would *degrade* performance because the potential for them to have overlapping schema sets is high. Conversely, similarities between to-be-remembered words and other to-be-remembered words within the same primary task would *enhance* performance, because their 'recall-list' schema overlaps with their semantic/categorical schema. Conlin, Gathercole, and Adams (2005) and Li (1999) both found evidence for the *degrade* in recall when there were similarities between the recall list and the distractor component, while there is an abundance of evidence that same-list similarity does indeed enhance recall (e.g., Poirier & Saint-Aubin, 1995; Saint-Aubin, Ouellette, & Poirier, 2005), evidenced through clustering effects (Bousfield, 1953; Manning & Kahana, 2012). Interestingly, Crowder (1979) also found the faciliatory effect of semantic similarity between elements on overall item recall, but a detrimental effect on *correct order* recall. Crowder's diverging result supports the idea of two distinct relational structures being established: a serial order constructed at the time of the task and a semantic similarity schema. The overlapping of the two relational structures improves unordered recall (because each set provides a method for the elements to be maintained) but competes on ordered recall (because only the serial order holds information on the sequence order). Saint-Aubin and Poirier (1999) however, found no disadvantage for semantic similarity on order accuracy. Thus, while the enhancing effect of same-list similarity is ubiquitous, a detrimental effect on order effects is more contentious. Oberauer (2009b) suggests that the conflicting different-list findings are the result of a trade-off between beneficial and detrimental propagation of similarity. The distractors add additional cues for recall, but also overwrite the features of the target list. Oberauer (2009b) found that different-list similarity only led to a detriment in phonological overlap (rather than semantic overlap). However, because Oberauer's (2009b) secondary task was based on pronouncing the distractors aloud, it is likely that the exclusive phonological overlap deficit was a result of task format. Thus, while the past research is overall consistent with a relational system of WM where different relations (e.g., temporal or semantic) can be applied to the same recall lists, it is important to consider task factors that may promote or obstruct certain types of relations. In this case, the semantic similarity of the stimuli was critically important, as was how the dependent variable was scored (e.g., whether order is critical to accuracy or recall can be unordered, and whether intrusions are ignored or penalized).

The current project makes the following assumption: relational integration is not simply a subcomponent process of WM but rather, it is the foundation of WM. Recall the three classic functions of WM: storage, processing, and attention. Now, consider how they manifest in a relational integration system of WM. In order for elements to be *stored* within WM, they must each be bound to a role (or, according to Oberauer (2009a), a content node is bound to a context node). *Processing* consists entirely of establishing and dissolving bindings, thereby integrating and disintegrating relations. Processing thus plays a part in storage (binding elements to roles in the first place) but also controls manipulation and generation of new relations (switching the bindings around to instantiate new relations). Already, the strict demarcation of *storage* and *processing* becomes nebulous. *Attention* is "taking possession by the mind" (Cowan, 1995, p.4): allowing a single content-context binding in WM to be focused by the conscious mind. Direct access is provided to multiple additional bindings, allowing attention to be redirected to any of these bindings in WM instantly and without retrieval from LTM. Attention therefore also relates to how bindings are stored and the depth of processing that can be accomplished. In this way, the three classic functions of WM (storage, processing, attention) do not together form a multi-componential WM but rather, they are emergent properties resulting from WM being based on relational integration. From this view, WM cannot be considered modular with a distinct 'operator' executing processes on modular systems (Baddeley, 2000). Rather, WM can only be segmented by the level of attentional activation provided (usually none for LTM, high for the direct access region of WM, and completely activated for the binding in central attention), and the system is dynamic rather than static. From the relational integration framework, WM still constitutes critical functions: attention, storage, and access to STM and LTM. But the lines between these functions are not as strict, because they all work in tandem, as they should.

Traditionally, the goal in WM research has been to explore differences in ‘storage capacity’ (the number of elements that can be maintained over time) in the face of processing disruption. This is a (superficially) useful metric because the measurement is easy to understand: the number of items that can be recalled is your capacity, and if you can recall more than another person, you have a higher capacity. In actuality, it has probably served to oversimplify capacity limits and led to casual reductionism. The storage-while-processing metric emerged from separate-component definitions of WM (demarcating storage and processing; Cowan, 2017), where the gold standard of measurement is the *complex span task* (Daneman & Carpenter, 1980; Engle, Cantor, & Carullo, 1992) which appears reliable across task domains (Conway et al., 2005). Complex span tasks require iterative retention of elements (the primary task) while alternating with some distracting processing task (the secondary task). For instance, the *operation span* variant of the task may involve alternating between remembering a word (like *CAT*) then solving a simple arithmetic equation (like 3 + 5 = ?), repeating this pattern until participants are probed to recall the list of words in the order they appeared. Although constructing a recall list itself requires updating a continuous sequence of bindings (of elements to serial position roles), the focus of this recall list as an indicator of capacity may be somewhat misleading. It may not be the capacity itself (in terms of number of raw elements) that defines individual differences, but how rapidly, accurately, flexibly, and firmly the bindings can be generated and maintained in the face of the repeated tangential processing (Oberauer & Hein, 2012). It is also important to consider how effectively the participants can chunk the individual raw elements to overcome the extreme number of elements (Cowan, 2001), which is directly related to all of these ‘ingredients’ of successful relational integration, such as flexibility. The proposal that distinct, intermittent processing in complex-span tasks ensures measurement of a true capacity is unlikely at best, and psychometrically indefensible at worst. Mathy, Chekaf, and Cowan (2018) demonstrated that the ‘chunkability’ of elements improved performance to a similar extent in both simple span and complex span tasks, indicating that the addition of processing tasks do not relate to a disruption in chunking. Unsworth and Engle (2006) suggest that the intermittent task simply raises the chance of displacing an item (accidentally unbinding a memory element). This may be related to intermittent tasks ensuring that participants cannot use sensory memory through subvocalization to rehearse the series of elements. In any case, chunking is almost assured to occur in complex span tasks, and complex spans probably only work as well as they do (in measuring WM capacity) because they, to a degree, measure chunking ability (Chekaf, Gauvrit, Guida, & Mathy, 2015), which can be more directly (through a relational integration perspective) seen as binding effectiveness. A major goal of this thesis is to verify this perspective of relational integration and demonstrate that understanding WM through a demarcation between storage and processing is a misleading and unsatisfying interpretation.

Before outlining these aims in more detail (along with the scope of the thesis), it is necessary to address the ever-present links between WM and higher-order “fluid intelligence” tasks that involve abstract reasoning (Ackerman et al., 2005). Fluid intelligence is general, flexible functioning applied to novel problem-solving situations, and has been described as the ability to rapidly learn, apply brain power, and extract information and patterns from complex set of stimuli (Blair, 2006; Cattell & Horn, 1978). Fluid intelligence is the hallmark of individual differences research, because tests of fluid intelligence have shown outstanding predictive ability for academic achievements (Giofrè, Borella, & Mammarella, 2017; Laidra, Pullmann, & Allik, 2007; Matešić, 2000) and workplace performance (Schmidt & Hunter, 2004), at least for modern, developed Western societies (Wicherts, Dolan, Carlson, & van der Maas, 2010). As it turns out, fluid intelligence tasks always involve working with relations, drawing on patterns and integrating relational information to uncover a missing piece of the puzzle (this is explored in Chapter II). Thus, where the classic modular view of WM is frequently at odds in attempting to correlate disparate tasks (simple ones that measure raw storage capacity and complex ones that measure abstract reasoning), a relational integration perspective of WM uncoincidentally suggests that similar processes act on both WM tasks and fluid intelligence tasks – this thesis will refer to this theory as the *relational integration hypothesis*. The relational integration hypothesis circumvents correlational arguments that see fluid intelligence as some ethereal, immeasurable attribute of humanity ("traditional views", as pointed out by Shipstead et al., 2016). For the relational integration hypothesis, the correlation is sensible. Although this thesis will continue to use the term *fluid intelligence* (Gf), the definitions between Gf, abstract reasoning, and even WM (from a relational integration perspective) may largely overlap (Deary, 2003), as all fundamentally demand relational integration. Chapter II outlines some key differences between WM and Gf. Although this thesis ultimately concedes that there are more similarities than differences, these differences have often been nebulously described due to an overreliance on individual difference theories, which cannot explain the processes underlying Gf (Birney, Beckmann, & Beckmann, 2019). The main point to take away from this short briefing on Gf is that factors associated with effective relational integration jointly influence performance on both WM and Gf tasks.

1.3. Aims of the thesis

Section 1.2. argued that WM should be understood as a system that allows for temporary binding between two types of mental representations: ‘content’ elements consisting of units of information, and flexible ‘context’ roles that signify positions within a relation. This is as opposed to a componential system with static stores which are operated on by a distinct central executive (Baddeley, 1992). Section 2.1. explores in more depth the issues related to componential views of WM, but the most pertinent issue is that the central executive is unfalsifiable and pervasive to understanding of a unitary WM system (Baddeley, 1998; Parkin, 1998). To rectify this long-standing issue, the goals of this thesis is to (i) demonstrate that WM can be understood most completely through a theory that incorporates relational integration, rather than through a system that demarcates storage and processing, (ii) explore factors associated with individual differences in effective relational integration (binding capacity manifesting through relational complexity, flexibility, and systematicity), and (iii) verify the relational integration hypothesis by demonstrating how relational integration jointly acts on WM and Gf tasks, and how it uncoincidentally accounts for the WM-Gf relationship over modular theories of WM which require more dubious explanations.

This project attempts to address these research questions with an experimental-differential approach, with each study involving either experimental manipulations or a combination of experimental and individual-differential techniques. In all cases (every study) and for every research question, evidence presented tends to reinforce theories of relational integration and weaken modular theories (and, at times, attentional control theories) of WM. The thesis should provide a compelling argument for why modular theories are unsatisfactory, and some evidence (Chapters V and VI) for why attentional control theories inadequately account for the overlap between WM and Gf. The remainder of this chapter is devoted to outlining how the current project attempts to achieve this goal by proposing the two core research questions addressed by this thesis: (i) How can binding capacity for relational integration be conceptualized and operationalized? (ii) Does the relational integration hypothesis account for the overlap between WM and Gf tasks?

1.3.1. How can binding capacity for Relational Integration be conceptualized and operationalized?

This project takes the theoretical position that WM in general can be best understood through measuring the capacity for relational integration rather than a capacity for storage or attentional control. Binding capacity can be thought of as the number of bindings that can be simultaneously active in WM. Support for a ‘binding capacity’ conceptualization of capacity limits has emerged directly from Oberauer et al. (2007) and Chuderski (2015), while research such as Chekaf et al. (2015) also support the theory but without the specific reference to a ‘binding’ capacity. The Relational Complexity metric (Halford et al., 1998) attempts to quantify the number of bindings involved in each relation, with the suggestion that tasks requiring a specific relational instance as an outcome can quantify the binding capacity required of that task. The Relational Complexity scheme also introduces the important concept of *systematicity:* how systematic a set of bindings are in a relation. Systematicity can be thought of as the ease of chunking but is more specific and can be identified at the task-level. High systematicity means bindings are consistently ordered in a fixed way, allowing problems which may appear to demand multiple active bindings to be solved with fewer active bindings (often only one) at the cost of some specific element-wise information that is often not related to the task solution. Throughout this thesis, we will see repeated examples of why systematicity must be considered when attempting to operationalize binding capacity.

Binding capacity is conceptualized distinctly from a more unspecified ‘storage’ capacity because it automatically accounts for chunking by the very nature of the model (rather than as an addendum to the model). It also does not assume the same connotations with active vs. passive storage (or ‘primary’ vs. ‘secondary’ memory). Bindings are only ever active because they appear in the direct access region of WM where attentional activation is high (see Chapter II for more detail). The delineation between active and passive memory does not need to be so strict, because LTM is virtually infinite, and because bindings can be expressly activated through recent activation or schemas. Thus, a binding capacity view circumvents many of the shortcomings of storage capacity and, in fact, we find repeated evidence throughout the thesis that passive storage is virtually irrelevant to task performance and in linking WM tasks to Gf.

1.3.2. Does the Relational Integration Hypothesis account for covariation in Working Memory and Fluid Intelligence tasks?

The relational integration hypothesis posits that binding capacity is the core limiting capacity that restricts performance on both WM and Gf tasks. As described earlier in this chapter, and explored in Chapter II, WM and Gf tasks overlap (at least in part) because they share relational integration demands. In four of the five studies (all except Chapter VII) presented in this thesis, prototypical measures of Gf are taken, and Gf performance is used as an outcome measure to demonstrate that the overlap between WM and Gf can be mostly (but not completely) accounted for by variation in relational integration demands, as predicted by the relational integration hypothesis. Differences in relational integration processes between WM and Gf emerge primarily due to the level of analysis. WM is often assessed at a level requiring maintenance and manipulation of bindings and the outcome of WM tasks is often verification that the binding is intact. Gf tasks, meanwhile, often require a comparison of two relations or the induction of an abstract rule that governs a set of relations. In either case, relational integration is fundamental, and a limiting factor of binding capacity affords only so much mental workspace for each participant. Because Gf tasks often have many complex demands (described further in Section 2.7), attempting to understand relational integration through higher-order Gf tasks may result in an unsatisfying level of doubt. Thus, our approach is to use operationally simpler tasks defined as ‘relational integration tasks’, which are (relatively) more process-pure, whereby the more variance that can be accounted for, the more the Gf tasks that are being predicted are actually based on relational integration.

It should be cautioned here that no measure is truly process-pure, so we cannot ever claim that relational integration can uniquely explain performance on Gf. This is because tasks will always require some degree of perceptual information extracted, some degree of mental representation, some degree of attention, goal maintenance, inhibition, and many other demands (see Section 2.7 for further discussion on this point). A failure of any of these demands will lead to an error and an error will (generally) be marked the same way on the outcome of a task. Thus, this aim is not so much an attempt to prove that relational integration is the most important be-all and end-all process of intelligence (a goal that would be doomed to fail) but instead, to simply highlight the consistently remarkable ability for relational integration to explain the WM-Gf overlap over modular theories of WM, which struggle even more with process-purity. Wherever possible, tasks are designed in ways that errors of relational integration are exemplified over errors of other demands. For instance, putting time limits on task items means that errors are now possible through an increase in failure conditions contingent on demands like processing speed, mind-wandering, and goal-neglect. While all of these demands may well still contribute to accuracy errors we observe, the tasks are designed to make this a less likely outcome (especially over many items). Thus, although we cannot ever truly claim that relational integration is the most important process, we simply pose the experiments as evidence of the relational integration hypothesis, and each reader may accept this as psychometric evidence differently. What should be clear though, is that, as a cognitive model of WM, relational integration is a critical demand for successful task performance and should be considered alongside more easily understood demands such as mental representation and focused attention.

An abundance of research has been devoted to attempting to understand the overlapping functions between WM and Gf, to the point where all theories of WM must be able to explain performance on Gf tasks also (Conway et al., 2007). At the higher-order Gf level, the specific task attributes are lost (or at best, obscured) either simply in the complexity of the task or in the condensing of tasks into a latent factor. However, at times, WM theories can still be difficult to disentangle even at the WM task level. For instance, this chapter earlier explained how a relational integration WM can explain complex span task performance. However, the standard complex span tasks do not offer a way to distinguish theories of attentional control from relational integration and in fact, complex span tasks are the hallmark paradigm of attentional control theories (Unsworth & Engle, 2006). Variation in performance on complex span tasks could contribute to either theory. In the current thesis, task analyses are conducted in each chapter, aiming to determine the relational integration demands involved in the core task. However, there are times when other theories (e.g., attentional control) may also explain the same (or similar) variance in the task, and thus, may also explain the WM-Gf overlap seen in that chapter. In general, although each individual study may not conclusively settle on relational integration as the core overlap (because it can be difficult to rule out attentional control theories), the thesis as a whole should provide compelling evidence for a relational integration view of WM and, at the very least, demonstrate how a relational integration view provides an explanation of the WM-Gf overlap that may not necessarily be conclusive, but is parsimonious.

1.3.3. Scope of the thesis

In concluding the introduction, it is important to briefly acknowledge important aspects of the research that do not fit within the scope of this thesis. These include LTM and the latent variable analysis.

1.3.3.1. Long-term memory and the procedural/declarative distinction

Although the current research focuses on WM, the important role of LTM and its place within this thesis must be addressed. In agreement with similar theorizing (Cowan, 2005; Engle, Kane, & Tuholiski, 1999; Logie, 1996; Oberauer, 2009a; Ruchkin, Grafman, Cameron, & Berndt, 2003; van der Linden, 1998), the current view is that WM is a subsystem or activated portion of LTM, rather than separate systems. The relevance of LTM to WM is apparent both from the use of schema sets (discussed earlier) but also to any task that involves passive storage-over-time where the contents are stored beyond a directly accessible state. In the current research, reference to LTM will primarily be made through these two avenues (schema sets and passive storage). However, LTM is also typically associated with a distinction between *declarative* (knowing 'what') and *procedural* (knowing 'how') memory systems (Squire, 2004). In an 'activated LTM' view of WM, this distinction persists at the level of active bindings. Oberauer (2009a) proposes two systems of WM (one declarative, one procedural) which operate in parallel: the declarative WM selects elements for operations and the procedural WM acts on them. Because both operate in parallel, it is often difficult to distinguish them at the task level. In general, declarative WM is more likely to be tapped in a task requiring binding and rearranging elements while procedural WM is more likely to be tapped in dual-task (task-switching) paradigms where goal orientation is essential (Kane & Engle, 2003). For the purposes of limiting scope, the current project does not distinguish declarative and procedural WM, though we acknowledge the usefulness of demarcating the systems at the WM level (Oberauer, 2009a).

1.3.3.2. Latent variable analysis and Gf

The second point to consider is the use of latent variable analysis. Latent variable analysis is useful because it extracts the commonality between tasks. Thus, rather than being concerned over task-specific artefacts or subtle differences in task presentation, the latent variable only accounts for what is common to all the tasks. This is particularly useful when measuring Gf, because we can supposedly capture more meaningful variance in Gf, rather than also picking up task-specific noise. The current gold standard for intelligence research is to also use latent variable analysis to extract a factor from WM tasks (Kane et al., 2004). The WM factor and the Gf factor are then correlated. In general, latent variable analysis of this kind (correlating latent variables to latent variables) can be unsatisfying. We are left considering what is common to each of the tasks which is often a variety of possible interpretations, from attentional control to storage. As we will see in this thesis, apparently similar tasks can have considerably different reasons for their variance (consider the discussion of the Arithmetic Chain Task compared to the Swaps in Section 6.4). Latent variable analysis also requires large sample sizes (*n* = 200+) and long testing sessions with many tasks which is arduous and impractical for this project which favours multiple studies. Instead, our preferred approach is to experimentally manipulate theoretically simpler tasks such as those highlighted in each study, to demonstrate the usefulness of task-level analyses. The different conditions of the task are thus identical in every way except for the experimental manipulation, allowing us to conclude that any differences seen in the variation in performance on the task (and how it relates to Gf) must be due to this manipulation. Although this solution leads to elegant conclusions, this approach is more susceptible to task-specific limitations. Thus, in every chapter, a task analysis is conducted to identify what demands are (or are not) associated with each condition.

Although our overall preference is against a purely latent variable approach, it must be acknowledged that there are benefits of a latent outcome variable such as Gf, to smooth out task-specific differences in applying the results to real world applications (e.g., Schmidt & Hunter, 2004). However, unfortunately, this was not always possible for every study, so the mileage of the Gf latent variables does vary between studies. In the worst case, there is only a single measure representing Gf. In these cases, we use an abbreviated Raven’s Advanced Progressive Matrices (J. Raven, 1989), as it is the most well-established Gf measure in the literature (Carpenter, Just, & Shell, 1990; Jensen, 1980).

II. THEORIES OF WORKING MEMORY

In this chapter, the theories most relevant to the current project are outlined. I preface this chapter with the admission that this review cannot be close to comprehensive, given the scale of definitions in the field (Cowan, 2017). Thus, the theories chosen for discussion here are ones that are most frequently drawn into discussion within this thesis. This chapter begins by describing the most popular model of WM, Baddeley and Hitch's (1974) multi-componential view, discussing its limitations and pervasiveness in the field. I then describe several additional models relevant to the current research (attentional control, generic working memory, the concentric model, and relational complexity) and explain how these provide more useful conceptualisations of WM. The chapter concludes with a synthesis of the relational theories (the models of Cowan, 2001; Halford et al. 1998; and Oberauer, 2009a) to develop a framework for assessing relational integration that is distinct from higher-order fluid intelligence.

2.1. The Multicomponential Model

In 1974, Baddeley and Hitch proposed a model of WM which combined two short-term storage systems (a visual store and an auditory store) with attention and processing, distinguishing WM from the more passive storage of STM (Atkinson & Shiffrin, 1968). The short-term storage systems were split into the *visuospatial sketchpad* and the *phonological loop*, following evidence that indicated somewhat distinct capacities for visual and auditory domains. Attention and processing functions were relegated to the *central executive,* a system for controlling allocation of resources and adjusting stored elements. Although the demarcation of visual and auditory domains was useful for addressing discrepancies between the modalities (e.g., Murdock Jr., 1966), the central executive was the catalyst that distinguished the *working* from the *memory*. Despite this, Baddeley (2003) concedes that the central executive is the least understood part of his model, resigned to descriptions of an "inscrutable, immaterial, omnipresent homunculus" (Donald, 1991, p. 327) making decisions on where to allocate attention and how to manipulate stored information. Baddeley (1998) has defended his use of the homunculus as a starting point: if we know that functions like attention and manipulation exist and are separable from storage, then we can conceptualise a system that handles these functions even if we cannot assign it to a neural substrate. However, as Donald (1991) and Parkin (1998) point out, the central executive has more often been relegated to handle functions that memory theorists cannot explain. The easily-understood homunculus analogy (see Figure 2.1) has potentially contributed to the enduring and widespread use of Baddeley and Hitch's multicomponential model both within WM (Cowan, 2017) and beyond, featuring in seminal visual perception papers (Luck & Vogel, 1997) and large reviews in neuroscience (D'Esposito & Postle, 2015), further propagating its status as the most influential view of WM.

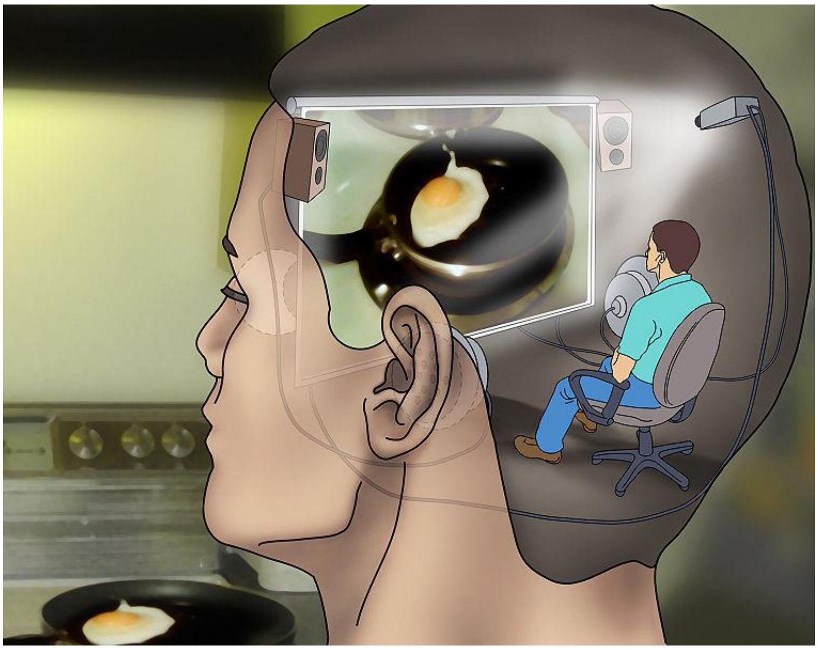


Figure 2.1. The central executive is often seen as a homunculus: a miniature "person inside your head" that dictates the focus of attention and manipulates contents of the storage systems. While the homunculus can be a helpful analogy, it has proved pervasive, warping common understanding of working memory. Image source: *Cartesian Theatre* (2008) by Jennifer Garcia. Reproduced with permission by CC-By-SA-2.5.

The failure to explain the central executive was largely a result of the times (Cowan, 2017) where passive STM systems were commonplace despite their inadequacy to explain processing (Atkinson & Shiffrin, 1968). Thus, Baddeley and Hitch can be commended for their influence on memory as an active system allowing for complex cognition. It is unfortunate that so much of Baddeley's attempts to clear misconceptions (e.g., Baddeley, 1998; Baddeley, 2012) about the central executive have not been as influential outside the WM field, though Baddeley (2000) did make a well-known amendment to the model by adding the *episodic buffer*. This component acted as a link between attention, LTM, and the slave storage systems, binding information together and allowing for the integration of elements in a similar way to relational theories. The specifics of how this binding occurs in the episodic buffer is not clear (like specifics of the central executive), but the fact that Baddeley felt it necessary is a clear indication of the pressing nature of relational processes in WM. Once again, this may have unfortunately set back common understanding of WM, as relational integration processes could be relegated to the episodic buffer without proper specification of the processes (once again, reminiscent of what occurred with the central executive). This is further setback by the misnomer 'episodic', inspired by Tulving (1983), despite the buffer's capability of integrating semantic representations. Although necessary for the multicomponential model to account for relational integration, the general ambiguity of the episodic buffer (Cowan, 2017) has resulted in another element of confusion for those trying to understand WM. This is particularly problematic considering the pervasiveness of the multicomponential model.

While Baddeley and Hitch's model was critical to extending WM beyond passive storage, the ambiguities of the central executive (for attention and processing) and the episodic buffer (for relational integration) have proven consistent issues. The episodic buffer, while a necessary addition, was a band aid solution. Although the multicomponential model explains the functions and limitations of the storage systems, contemporary theories have begun to demonstrate that central executive functions (attention, manipulation, integration) explain more variance in WM capacity than storage aspects (Halford et al., 1998; Oberauer et al., 2008; Shipstead et al., 2016), so we cannot rely on a system that only details storage. Although there have been other models focusing on storage aspects of WM (Colom, Shih, Flores-Mendoza, & Quiroga, 2006), the primary reason for considering other models beyond Baddeley and Hitch's is the ambiguity surrounding the central executive and the episodic buffer. Therefore, the remaining theories considered in this chapter tend to focus more on processing.

2.2. The Attentional Control Model

Emerging from the insufficiency of Baddeley's model to explicate the central executive, Engle and colleagues conducted research highlighting the importance of controlled attention (Engle & Kane, 2004; Engle, Kane, et al., 1999; Kane, Bleckley, Conway, & Engle, 2001). Engle (2002) suggests that WM capacity is not based on storage capacity (in the sense of the number of elements that can be stored) but is instead based on the ability to control attention, focusing on task-relevant elements while preventing proactive interference from both on-task information (e.g., elements already recalled) and off-task information (e.g., 'what's for lunch?'). In this way, WM capacity is measured just as much through recall of a single element as it is through recall of seven elements. According to this view, complex span tasks (Daneman & Carpenter, 1980) are not good assessments of WM because they 'combine storage and processing' (the two functions of WM into the primary and secondary tasks) but rather, because they load highly on the ability to sustain and shift attention (between the primary and secondary task). Similarly, a controlled attention view accounts for dual task performance where those 'low' in the ability to sustain attention are more easily distracted (Colflesh & Conway, 2007; Conway, Cowan, & Bunting, 2001).

At times, the attentional control approach to experimental research in WM is to divide participants into low and high WM capacity (Engle, 2018) based on performance on WM tasks (usually complex span) – lower and upper quartiles representing ‘low’ and ‘high’ capacity individuals, respectively. ‘Low’ participants demonstrate numerous difficulties in controlling attention compared to ‘high’ participants (Engle, 2002). The issue with this approach is that the cut-offs are arbitrary and implies that the researcher is confident both that the measures (complex span tasks) are measuring the full range of abilities, and that the cut-offs (quartiles in this case) represent a qualitative difference, despite often using highly educated university student samples. Engle and Kane (2004) acknowledge this limitation of the extreme-group paradigm, and yet Engle (2018) still patronizingly refers to the ‘low’ WM group as being distracted by “pretty butterflies” (p.191). Attentional control theorists have also used latent variable analysis for more comprehensive research into individual differences, with Engle, Tuholiski, Laughlin, and Conway (1999) finding attention control critical for WM to predict fluid intelligence, while STM alone could not (but see Colom, Shih, et al., 2006). Findings such as this makes it no wonder that researchers are questioning the term *working memory* in favour of the term *working attention* (Baddeley, 1993).

The premise of attention control theories appears more elegant than multicomponential theories because all variance boils down to the use of executive attention, rather than storage capacity with some unknown contribution of the central executive. However, this does raise additional questions about (a) how information is stored in WM and (b) the processes used to direct attention. Unsworth and Engle (2007a) propose a distinction between *primary* and *secondary* memory. Primary memory consists of the elements stored in central attention, while secondary memory consists of activated elements temporarily kept outside the focus of attention. For instance, during a complex span task, to-be-remembered elements are encoded into primary memory but must be quickly displaced to secondary memory in order to deal with the processing task (Unsworth, Spillers, & Brewer, 2010). The concept of primary and secondary memory is analogous to the distinction between active and passive storage, which is prominent in both the following relational WM theories (Cowan's and Oberauer's) and thus, the preferred terminology in this thesis (used primarily in Chapter IV). In terms of redirecting attention, Shipstead et al. (2016) propose two functions: *maintenance* of the currently attended element and *disengagement* from irrelevant information, with engagement of attention through attentional capture or top-down executive signals. Shipstead et al. acknowledge the similarity to a relational integration binding system (Oberauer, 2002): binding, maintaining, and unbinding reflective of engagement, maintenance, and disengagement.

The controlled attention view has made substantial progress in distinguishing working memory from the multicomponential system, detailing the role of executive functions such as attention. Engle's (2002) paper reviewing evidence for controlled attention was seminal. However, like Baddeley's influence, Engle's review has spread beyond cognitive psychology and, like Baddeley's homunculus analogy, this may be because it was written to be easily understood by a wide audience. It is unfortunate to see that history may be repeating itself as (also like Baddeley) the attention control approach may have become pervasive, with the theory's defining paradigm (the complex span) frequently producing statistical inconsistencies such as correlating more with short-term memory measures than WM measures (Colom, Rebollo, et al., 2006) and poor correlations to other supposed WM tasks like the *n*-back (Redick & Lindsey, 2013). Engle's (2002) review also sparked the belief that WM capacity is thoroughly distinct from, and can predict, fluid intelligence (Gf). Although there is a longstanding, consistent, and powerful relationship found between WM and Gf (Ackerman et al., 2005), this belief implies that the WM and Gf stand as independent constructs. As Shipstead et al. (2016) state, this view treats WM as "something concrete and elemental, while fluid intelligence remains a divine outcome" (p.772). Engle (2018) concedes that it was wrong to predict Gf through WM tasks as in his 2002 review, now agreeing that performance on both Gf and WM tasks come about through similar functions. For Engle and colleagues, this means attentional control. Despite the considerable overlap between the attentional control model and the relational models (presented later in this chapter) in their predictions, the two perspectives often have only a modest amount to say on one-another (Cowan, 2017; Shipstead et al., 2016). Because the primary perspective of this thesis is of a relational WM, it is unfortunately likely that this thesis will also underrepresent attentional control theories. Wherever feasible, differences between the two perspectives are identified and contrasted (see Chapters III, V, and particularly, VI) but it should be concluded (once more) that the two perspectives are more similar than they are different.

2.3. Cowan's Model of Generic Working Memory and theory for chunking

Cowan (1988) maintains that working memory is ultimately about temporarily accessing a limited amount of information. Similar to Baddeley (2000), generic views of WM (Cowan, 2017) tend to be agnostic towards processing aspects, stating that WM must have a repertoire of functions for processing information but we do not yet know enough to explicate these beyond a central executive. Unlike modular views, a generic view does not see WM as a distinct system with distinct subsystems (e.g., an auditory store and a visual store). Rather, WM is the activated portion of LTM (embedded within LTM, see Figure 2.2) (Cowan, 1988), where information is temporarily accessible to executive processing (attention or manipulation). In this way, WM and LTM cannot be functionally separated, and it may be better to consider the system simply as 'memory', while terms like WM and LTM only help to delineate levels of activation within 'memory'. Cowan (2001) highlights this by noting that attention to exogenous information still involves activating representations in LTM – listening to another person speak does not involve representing the acoustic waveforms in

WM but rather, they activate the meaningful symbolic representations connected to the sounds from LTM (Cowan, Winkler, Teder, & Näätänen, 1993).

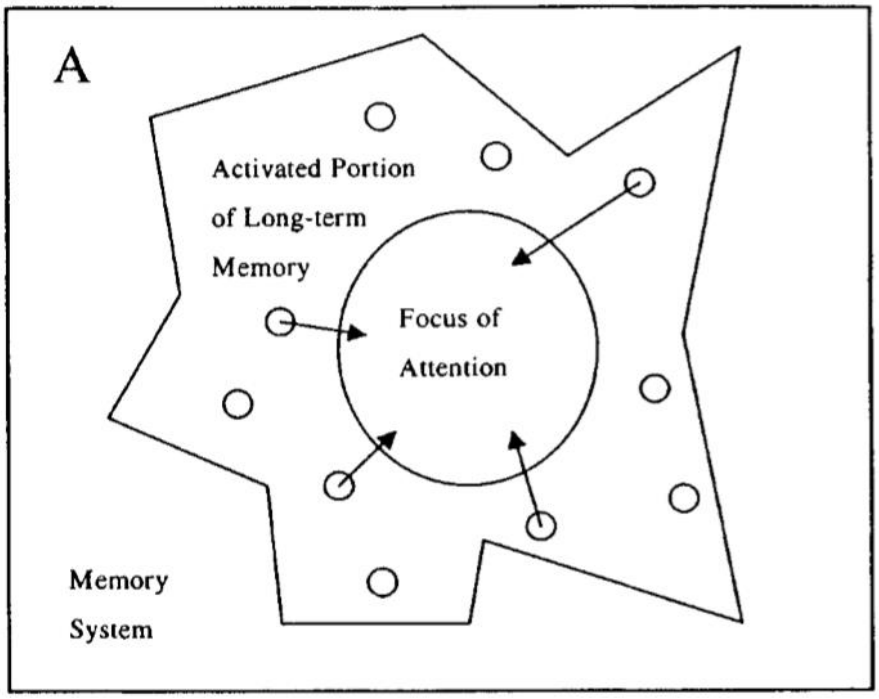


Figure 2.2. Diagram from Cowan (2001), representing WM as an activated portion of LTM. This contrasts with models of WM from Atkinson and Schiffrin (1968) and Baddeley (2000), where WM and LTM are distinct components.

Generic views of WM are cautious about dividing functions through experimental task manipulations. For instance, whether information is lost as a result of decay or interference is a difficult question to answer empirically (Cowan, 2001), though there has been success with computational modelling (Lewandowsky et al., 2010). Similarly, it makes little sense to experimentally divide processing and storage components (Daneman & Carpenter, 1980) because even the most basic storage tasks involve activating WM and forming an ordered list. This can explain why researchers find success in measuring WM using tasks without a clear processing aspect (Colom, Shih, et al., 2006) or without a clear storage aspect (Bateman, 2015; Chuderski, 2014; Oberauer et al., 2008). The current research hopefully makes it clear that the storage aspect being ‘experimentally removed’ is specifically passive storage-over-time (stored outside the focus of attention), as a general storage aspect cannot ever truly be removed from WM (activated representations must be stored active).

Cowan’s most significant contribution was his refinement of Miller’s (1956) chunking theory. Consistent with an inexact, generic view of WM, Cowan (2001) defined a *chunk* simply as “a collection of concepts that have strong associations to one another and much weaker associations to other chunks concurrently in use” (p. 89). This broad definition allowed for dynamic chunking (e.g., *cat* could be linked more closely to *dog* than *tiger* through ‘domesticated’, or more closely to *tiger* than *dog* through ‘feline’, depending on which association was active) and aligned with an ‘activated LTM’ view of WM. Cowan proposes that chunks are the base unit of measurement for capacity limits (rather than individual units or elements). Remembering a list of words is not about encoding each word as a new representation into memory. Rather, it is about activating words that are already established in LTM and temporarily generating associations[3] between them to assist in the simultaneous activation of many words beyond the limits of activation. Thus, storage capacity is based on how many associations can be constructed and held.

Cowan believed it was unfeasible to get a truly accurate measure of storage capacity, because chunks cannot always be identified (i.e., we cannot reliably identify how many elements or even what elements are part of each chunk). Instead, he proposed a distinction between estimates of capacity limits (number of chunks) and compound estimates (number of elements), then outlined steps that could be taken to identify chunks (and thus, the base unit of measurement for capacity limits). This proposal meant that when chunks could not be identified, compound estimates could be employed as a more accurate representation of the operational measurement (e.g., the number of to-be-remembered objects presented) with the

[3] Note that Cowan’s *associations* resemble the *relations* discussed in Chapter I. For this section, Cowan’s preferred term will continue to be used but outside of this section, relations will be the preferred term.

trade-off that the ‘true’ capacity in terms of chunks was unknown. Cowan’s methods of identifying chunks included overloading information (diverting attention), blocking recoding of elements (preventing rehearsal), and discontinuous performance at certain levels (e.g., marked drops in performance for enumerating more than four objects, but not less). He also suggested that established associations (from LTM) should be strong within the chunk and weak between chunks, encouraging participants to chunk in a certain way. For instance, it is far easier to identify the chunks being employed in a recall list like *cat-tiger-leopard-van-plane* than it is to identify the chunks in *cat-dog-tiger-mouse-leopard*. The former sequence clearly delineates two associations (animals and vehicles) while the latter involves several overlapping associations (animals, felines, domestic, wild). In a recall task, we would expect most participants to follow the well-established associations in the first sequence, while participants vary in how they chunk the second sequence, meaning we cannot identify the chunk capacity limit and cannot compare individuals by “capacity limits”. Similar clear delineation could be achieved by exogenous cues like Gestalt grouping principles or punctuation (McLean & Gregg, 1967).

Cowan’s (2001) substantial review of the literature led to the conclusion that the average capacity of WM is four chunks. This was made with several provisos. For one, Cowan suggested that chunks could hold a theoretically endless amount of information by incrementing on the associations to include more information. Asking for the free recall of any words associated with the word *cat* could result in dozens or even hundreds of associated words being recalled through long-term associations but asking for the specific recall of words only presented in a span task is probing a chunk that was only generated at the time of the task. This is like the system described in Section 1.2, where associated words like *doctor-nurse-hospital* could be recalled using either a novel relation (serial order from the task) or recall using the well-established ‘medical’ relation. Another proviso was the related strategy of elaboration (Cowan, 2001): where no well-established association exists, associations can be recoded using elaborative rehearsal, forming images that associate elements in some way. This is the key behind mnemonic systems allowing the recall of large sequences of elements (e.g., a deck of randomized cards) through elaboration of intermixed episodic associations (Bower, 1970). The final and crucial proviso is that the chunk capacity limit (of four) is a limit in the focus of attention, rather than WM *per se*. Many (if not infinite) representations can be passively activated, but only a maximum of four (±~1) can be active within the focus of attention. In this way, Cowan's model aligns with controlled-attention models in that ultimately, capacity is based on limits of attention rather than passive (secondary) storage. This approach also helped Cowan (2001) to suggest a teleological account of capacity limits, as a limited focus of attention would benefit search procedures (through hierarchical organization of elements) and comparative judgements (differences are exaggerated when using small sample sizes, compared to large sample sizes). This proviso also forms one of the only differences between Cowan's model and Oberauer's model (featured in Section 2.4): Oberauer (2009a) see the focus of attention limited to only a single binding but with a limited region of direct access accounting for the remaining active bindings.

Compared to multicompential models, Cowan's (2001) model simplified memory into one system based on activation rather than storage capacities, being careful to delineate what we know and what we do not. Cowan also provided a framework for chunking that illustrated how associations between elements were ultimately the basis of capacity limits, and how they could be measured. It remains a general model of WM, careful not to explicate processing. The next two sections outline models that aim to conceptualize processing and tie it all together to explain WM completely.

2.4. Oberauer's Concentric Model

Because Section 1.2 provided an outline of Oberauer's (2009a) model, this section is devoted to additional theoretical details. As a reminder of Section 1.2, Oberauer (2009a) states that representations in memory exist in two types of nodes: *content* nodes (including representations of objects, words, or events, such as elements in a series) and *context* nodes (including representations of roles or positions, such as place within a series). Content nodes are bound to context nodes to signify an element's role in the relation. Together, a set of content-context bindings constitute a relation with each binding representing one argument in the set. Instantiating and comprehending a relation thus means that the bindings have been integrated. Because any content node can be bound to any context node, an infinite combination of relational structures can be constructed, with the only limitation being the number of bindings that can be held active at any one time.

Like Cowan's model and the attentional control view, Oberauer's concentric model (Oberauer, 2002, 2009a; Oberauer et al., 2007) centres on attention providing temporary access to memory. The concentric model divides memory into tiers based on the level of attentional activation. Figure 2.3 demonstrates this model visually (Oberauer, 2009a). The highest level of activation is the focus of attention, where a single binding (i.e., a tethered content-context dyad) is held in immediate conscious awareness. The binding in the focus of attention is the strongest, shielded from interference by conscious attention. It is also the most flexible, as binding and unbinding of these representations can occur most freely at this level.

The next level is the *region of direct access*. At this level, a set of related representations are activated above threshold, granting a privileged status where they are available to immediately be brought into the focus of attention when demanded. To be activated to this level, content representations are bound to contexts that signify a shared relation. Rather than decaying over time, elements are unbound when they are no longer related to the currently activated relation. This is because new, unrelated representations may be brought into the direct access region and cause attentional interference (Oberauer, Lewandowsky, Farrell, Jarrold, & Greaves, 2012), causing the focus of attention to struggle to maintain the most activated element. This shares Cowan's teleological argument for capacity limits. However, unlike Cowan's theory, the extra layer (the direct access region) means that unwanted representations *may* cause interference, but not necessarily *will* cause interference. Novel relations are also generated in this direct access region, with binding occurring serially on each element until a relation is constructed. In a problem-solving situation, this novel relation may be congruent with the problem solution (in an analogy task), or it may produce a novel element congruent with the problem solution (in most other tasks).

Below this level (outside the direction access region) are representations that have some activation above baseline. This could be because they were recently activated to a higher level or because they are implicitly associated with representations currently active in the direct access region. For instance, a *fire truck* representation in the direct access region may provide some activation above baseline to *firefighter*, *ambulance*, and *red* even though none of those elements have been within the direct access region or central attention recently. This associative activation allows for quicker and easier access (as in priming), though aside from this associative activation, there is nothing qualitatively separating these activated elements in LTM from elements in LTM with no activation. However, importantly, activated representations outside the direct access region are qualitatively distinguished from elements within the direct access region in that they are not bound to contexts (roles). Well-established relations propagate activation between common representations (content and/or contexts) but representations are not bound until they reach the region of direct access. This allows these above-baseline representations to still be recalled (as in a span task) through cued retrieval by associating activation of targets with cues, or as a list by cascading gradients of activation levels through the list items (Oberauer, 2009a). Crucially, these methods are not as reliable as preserving the original bindings because the associations may be too weak to spread activation, but they may be the only option when processing is significantly diverted (as in complex span tasks).

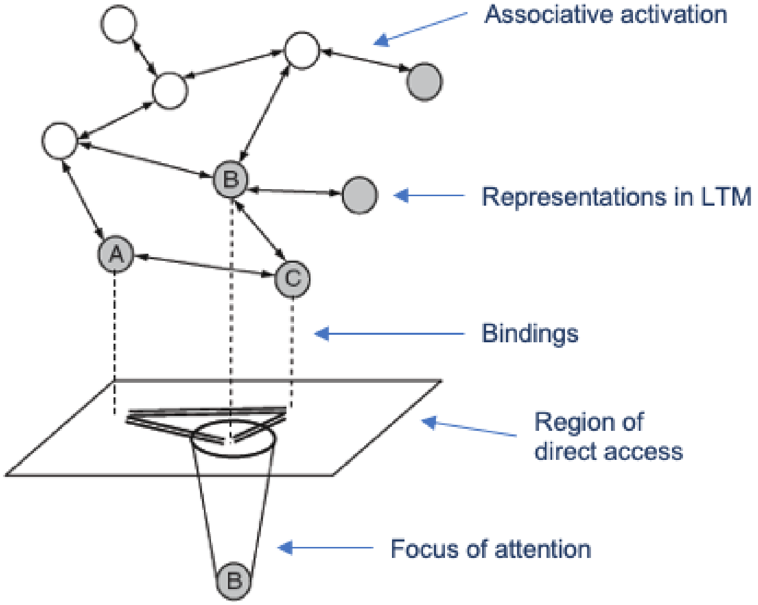


Figure 2.3. An *architecture of declarative working memory* from Oberauer (2009a), with labels added. Circles represent individual representations (/elements) in long-term memory and bidirectional arrows representing associations between representations. Currently, elements A, B, and C (represented within content nodes) are all bound (dashed lines) to roles within an interconnected relation (represented within context nodes; lined triangle) within the direct access region (the rectangular frame). Element B is currently within the focus of attention (cone). The bound elements propagate activation to associated elements, though these elements may or may not necessarily be activated above baseline (depicted by shading).

A final point to consider in Oberauer's model is dimensionality. Like the concept of dynamic chunking in Cowan's model, memory elements can be represented in several dimensions. For instance, *fire truck* could be represented on dimensions for physical space (next to us or far from us), hypothetical space (within a fire station or on a road), colour (red or white), as a category (emergency vehicles or land transport), or a theoretically infinite other number of dimensions. LTM stores associations on all these dimensions but when the element is represented in the direct access region (i.e., when it is bound), only a limited subset of these dimensions is activated. For instance, *fire truck* could be bound to the *vehicle* category in a relational instance of *drives(driver,vehicle)* and, although associations on many other dimensions are tangentially activated (e.g., red, ambulance), the dimension of the current binding promotes activation of elements that share the dimension of the active relation like *firefighter* to fill the *driver* role. One merit to our highly flexible generative memory system is that any element could fill the *driver* position, but it may require some creative and effortful thinking to construct and comprehend the relation. For instance, *dog* could fill the *driver* role, but it would not make much sense unless we also manipulate the dog to take on unnatural attributes like paws that can reach the pedals or work a gear shift. Dimensionality allows the memory system to be capable of generative thought while also promoting common, logical declarative thought.

2.5. Halford’s Relational Complexity Framework

The final theory to consider is Halford’s Relational Complexity (RC) model (Halford, Baker, McCredden, & Bain, 2005; Halford et al., 1998). This model differs from previous models in that it was not developed as a model of WM *per se* but rather, a framework for assessing processing aspects of WM. Despite this, it has been a strong influence in future models of WM (Cowan, 2001; Oberauer, 2009a). Where Baddeley’s model is insufficient in explaining the central executive, Halford designed RC as a metric for quantifying processing capacity. As it turns out, the pervasiveness of chunking in quantifying WM capacity means that RC may be a more appropriate method for measuring WM capacity in general, not just the processing features of WM.

In this framework, processing capacity is defined as the number of arguments that must be simultaneously represented to instantiate a relation between the arguments. Like similar theorizing (Hummel & Holyoak, 2001; Oberauer, 2009a), arguments are bound dyads (elements bound to roles, slots to fillers, contents to contexts). A relation of *binary* complexity would consist of two arguments. For instance, comprehending a relation of size between a rat and a mouse (RC=2, because there are two arguments) involves instantiating[4] a relation between *rat* and *mouse*, such as *rat-larger, mouse-smaller* allowing us to comprehend that the rat is larger than the mouse. Each additional level of RC involves an additional argument: *ternary* relations involve three arguments, *quaternary* relations involve four, and so on. *Unary* relations consisting of a single argument are also possible, but they only allow the isolated comprehension of an element's category (e.g., *rat-rodent*) or attribute (e.g., *firetruck-red*) which do not act as an operator to the relation.

Transitive inference problems are a clear way of showing how RC level can increase systematically. For instance, given the premises "John is taller than Mary" and "Mary is taller than Anne", we can construct relations using John, Mary, and Anne as elements and their heights as roles. Comparing John and Mary or Mary and Anne involves instantiating a binary relation because we need only consider one of the premises (two arguments): John is taller than Mary *or* Mary is taller than Anne (each are given in the premises) Conversely, comparing John and Anne would involve instantiating a *ternary* relation because both premises (all three arguments) must be simultaneously considered to deduce that John is taller than Anne: John is taller than Mary *and* Mary is taller than Anne, therefore John must also be taller than Anne. Once this ternary relation has been comprehended and we know that John is taller than Anne, we can compress the relationship between John and Anne into a simpler binary relation, with the proviso that this new binary relation cannot (on its own) offer information on how Mary fits into this equation.

[4] *Instantiating* a relation, *constructing* a relation, and *generating* a relation can be considered largely synonymous though there are subtle differences which dictates their use in this thesis. Halford prefers instantiation as a verb as it is similar to *representing* the relation in memory (you *instantiate* a relation as you would *represent* an element within WM), while Oberauer prefers the more processual term *constructing* which demonstrates that the relation is built through a set of bindings. Thus, this thesis will use 'instantiate' for the use of representing a relation in WM, and 'construction' for the more general building of a relation, while 'generation' will be preferred for times when a *novel* relation is constructed, such as for novel problem-solving.

Halford et al. (2005) present evidence that quaternary (RC=4) relations represent the typical upper limit of complexity that humans can process (uncoincidentally similar to Cowan's (2001) chunk limit of four). An important caveat to the RC framework is that the content of the elements or roles are distinct from the complexity of the relations, as is the format of the task. We could, for instance, make individual arguments more difficult to comprehend (e.g., by blurring the rat so it may be confused with the mouse, or by describing John's appearance rather than simply naming him) but this increase in difficulty would be on a separate scale to the difficulty being represented by RC. In this way, a task requiring binary integration could be more difficult (as in the likelihood of answering correctly) than one requiring quaternary integration, because there are factors independent on RC influencing the difficulty. Systematically increasing RC means keeping these other sources of difficulty constant across levels of complexity, else we introduce noise in the metric. In general, we cannot directly compare relations of equal complexity across tasks because there are many other factors that play into the difficulty of a task. For instance, solving the simple arithmetic problem *3+5=?* involves instantiating a ternary relation[5] between three arguments: *3-addend, 5-addend, ?-sum*, through which the simple incrementing of 3 to 5 results in 8, which can be retrieved and bound to the *sum* position for the solution. Conversely, a task like Raven matrices (J. Raven, 1989) also involves ternary relations but is unquestionably more difficult than one-digit arithmetic because, although the RC remains at three throughout the test, items range dramatically in difficulty due to the range of unknown rules (Carpenter et al., 1990) and the embedding of complexity in superimposed (3x3) sets of ternary relations (Birney,

[5] I acknowledge that one-digit addition likely involves immediate retrieval of solutions in adults due to over-learnt associations between single integers, but mental arithmetic is theoretically a ternary process (Halford et al., 1998). If this scenario seems unrealistic, consider the same example but with two-digit numbers.

2002). Thus, it is not advised to compare complexity across tasks but, with all else equal, increases in RC should equate to increases in binding capacity required.

2.6. Synthesizing the discussed theories

Five models of WM have been discussed in detail. Although each provide a perspective on WM, a synthesis is warranted to justify the understanding of WM used in this thesis. The vision of WM in this thesis is primarily based on Oberauer's concentric model, as it is the most comprehensive for investigating relational integration. While Halford's RC framework also applies, it was developed more as a metric for quantifying processing complexity in tasks (and will be used in this thesis as such). Cowan's model predicts similar outcomes to the concentric model but is purposely more general than Oberauer's. The attentional control view also predicts largely similar outcomes, but with a focus on the attentional aspects rather than the relational aspects of WM.

Overall, these models generally predict quite similar outcomes. Although several studies in this thesis attempt to distinguish relational theories from attentional control theories (in particular, Chapter VI), it may be disappointing to readers to see that the outcomes of the studies do not always conclusively rule one as superior. It is therefore important to point out again that the intention of this thesis is to demonstrate the usefulness of a relational integration approach to understanding WM, which has otherwise been underrepresented in the shadow of overlapping attentional theories. In Cowan's (2017) definitions of WM, attentional control models are separated only because they come from a large (and largely segmented) body of research that focuses on central attention; not necessarily because they disagree with generic definitions where WM is simply an activated portion of LTM. Recently, Shipstead et al. (2016) appreciated the usefulness of "temporary associations (bindings)" (p. 782) from Oberauer's (2002) concentric model to the attentional control view: flexible bindings between elements and a schema map create associations between elements (allowing for novel combinations). The bindings thus involve *disengagement* and *engagement* – central processes to the attentional control's perspective on capacity limits. Similarly, Cowan sees capacity limits in terms of chunk limits (associations) and Oberauer sees capacity limits in the number of bindings. All three models agree that WM capacity is best seen not as a restrictive capacity limit, but as a permissive buffer allowing only the most task-relevant information to be directly accessible.

Although binding is clearly critical to capacity limits for all models, there are some differences in how the ability to establish or dissolve bindings manifests in tasks and how it translates to higher-order cognition. This thesis aims to contribute to this understanding. While Cowan remains cautious about commenting on the specific processes involved, Shipstead et al. (2016) see the ability to effectively unbind (disengage) as the critical variation between individuals, tying back to attentional control through task switching and inhibition of distractors. Oberauer sees a more general capacity in the direct access region represented by the ability to combine all the necessary elements. Halford's RC metric relates closely to Oberauer's vision, in that the complexity of a relation may exceed the binding capacity of WM and be simply impossible to instantiate without severe compromises. Oberauer and Halford are mostly in agreement in that the complexity of relations contributes to binding capacity, but complexity rarely tells the whole story of relational integration. This was made clear by the example comparison of the two ternary tasks, simple arithmetic to Raven matrices, given in Section 2.5. Consider also, the task of making an analogy such as "A is to B, as C is to ?" (A::B=C::?). This task requires mapping the A::B relation onto the incomplete C::? relation. All four elements involved contributes to the capacity limits of the direct access region, though the effective RC remains only binary because both relations involve only two elements. As discussed in Section 2.5, there are a multitude of factors that contribute to performance apart from complexity, and these factors include the number of elements involved, not just the number of elements in the most complex relation. In concluding this section, it is worth quoting Oberauer's (2009a) rather simple definition for binding capacity: the ability to "put [all the relevant pieces] together by binding them into a common schema" (p. 92). Despite the subtle differences in the models pointed out throughout this section, this definition sets up the broad aim of this thesis: to explore the determinants that contribute to this capacity for binding. As Shipstead et al. (2016) state, Oberauer's (2002) model "bridges the gap" (p. 783) between perspectives, indicating that the unified understanding of WM that Baddeley (2012) is hopeful for may indeed be transpiring.

2.7. The relationship between Working Memory and Fluid Intelligence

Thus far, we have discussed the conceptual properties of WM and how binding processes can result in relational integration. During this discussion, occasional descriptions of tasks such as simple span, complex span, transitive inference, mental arithmetic, and Raven matrices have helped to illustrate a point. Before concluding this chapter, it is necessary to discuss how relational integration can be separated from other demands in tasks such as these, particularly in separating purely relational WM tasks from higher-order Gf tasks.

Much of the work on relational processing in cognition has been approached from a reasoning, intelligence, or mental abilities perspective (Dumas, Alexander, & Grossnickle, 2013), with educational research also seeing benefit in applying relational thinking to knowledge acquisition (Alexander, 2016; Resnick, Davatzes, Newcombe, & Shipley, 2016). While these approaches do not always discuss WM's contribution to relational processing, they have confirmed that abstract reasoning tasks involving multi-layered relations are the best assessments of fluid intelligence: matrix reasoning tasks such as Raven matrices (J. Raven, 1989) or Wechsler's (2008) matrix reasoning subtests are powerful predictors of general mental ability and scholastic achievement (Giofrè et al., 2017; Koenig, Frey, &

Detterman, 2008; Laidra et al., 2007; Matešić, 2000). Abstract reasoning tasks represent the current gold standard for measuring general fluid intelligence, and are thus often referred to as *Gf tasks* after Horn and Cattell's (1966) *general fluid (Gf) intelligence* classification. The literature has preferred the use of the term 'Gf tasks' (Ackerman et al., 2005) rather than 'abstract reasoning tasks', focusing on the novelty and non-verbal content of the tasks. The tasks are framed as culturally fair and intended to be administered to participants with no prior experience with the task. In reality, the evidence suggests that Gf tasks such as Raven matrices do draw on experience (Mervyn et al., 2002). In any case, given the high overlap between WM tasks and Gf tasks (Ackerman et al., 2005) and the contribution of relational integration to performance on both tasks, Gf represents a good platform to apply this research with ecological validity, even though evidence presented in this thesis indicates that relational WM tasks predict similar ecologically valid outcomes. From the conception of this thesis, the research was not yet there to circumvent Gf tasks entirely, but research by Birney, Bowman, Beckmann, and Seah (2012) and Krumm, Lipnevich, Schmidt-Atzert, and Bühner (2012) did presented initial evidence that relational WM tasks may be just as useful as Gf tasks for real-world applications such as assessments.

As briefly discussed in Chapter I, Gf tasks tend to involve multiple requirements, some of which can be isolated (Carpenter et al., 1990). For instance, perceptual information must be extracted, and a mental representation must be formed while non-relevant information must be inhibited. Attention must be sustained over time and goal-direction is necessary to orient and sustain attention over steps of a problem. Each of these demands are represented in both Gf and relational WM tasks. These demands are present in most cognitive tasks and a failure in any of them will lead to an error in the task. In addition, there are two additional demands that are relevant to this thesis: *rule induction* and *relational integration* (i.e., binding capacity). They are of interest because these two demands regularly reoccur as representing what is fundamentally different between WM and Gf tasks – rule induction; and what is fundamentally similar about them – relational integration. Although the prior demands (e.g., perceptual information, sustained attention) are also essential prerequisites of cognitive performance in both WM and Gf tasks, there is considerable evidence suggesting that the overlap with Gf can be explained by relational integration, and the gap to Gf can be explained through rule induction. The remainder of this section will highlight research contributing to this statement.

Rule induction refers to inducing the rules that govern a pattern between elements in a series. Carpenter et al. (1990) suggests that one of the most difficult aspects of Raven matrices is rule correspondence: identifying the rule involved in each matrix. Consistent with this, Verguts and De Boeck (2002) found that participants exposed to one particular type of Raven problem were significantly more likely to solve a sequential problem if it had the same solution rule as the prior problem, compared to a different solution rule. Bui and Birney (2014) extended this with their finding that participants only benefitted from repeated exposure to the rule when they correctly answered an earlier problem governed by the rule.

While there are a core set of rules governing Raven problems (Carpenter et al., 1990), the same rule can be represented with different surface features – using different shapes, for instance. Conversely, a different rule could be represented using a similar shape. Rule induction involves identifying the rule despite these surface variations. While Raven matrices tend to involve abstract shapes and patterns, a simpler example of the same principle is demonstrated by Vendetti, Wu, and Holyoak (2014) in their relational mapping task. Participants must recognise that an umbrella in one scene is analogous to a newspaper in another scene when they are both being used as shields from the rain, despite the second scene also containing an unrelated umbrella not being used as a shield. Gentner (1983) describes this as a difference in attribute mapping (umbrellas look similar) and relational mapping (the umbrella and newspaper are being used for the same purpose).

This research demonstrates that a significant portion of difficulty in Raven matrices comes from discovering the rules associated in a problem and knowing when to apply them. Thus, a crucial question to discern the relationship between WM and Gf is: what happens when the rules are given to participants? Loesche, Wiley, and Hasselhorn (2015) taught participants five of the core rules before taking Raven's Advanced Progressive Matrices. As expected, the authors found participants had significantly higher performance on the task compared to controls with no training. However, they also found that participants in the given-rule condition had *higher* correlations to WM measures than control participants taking the normal task. This seems to indicate that rule induction is an aspect of Raven matrices that does significantly impact on performance, but it is a demand *independent* from WM. Although the WM-Gf correlation did not rise to perfect in the given-rules condition, this does give reason to suspect that rule induction is a central unique component that separates Raven matrices from other WM tasks. When this is removed, the similarities between the tasks (i.e., both relying on relational integration processes) are amplified.

Additional support for this rule induction comes from the Latin-Square Task (LST; Birney et al., 2006). The LST is a matrix task that superficially appears similar to Gf tasks that also involve matrices. The LST is reminiscent of Raven's in that it consists of solving incomplete matrices by working out which shapes fill the empty target cells. However, the core difference is that the LST involves a single rule that is made explicit to participants in the instructions: *each row and each column may have only one of each shape*. Previous research with the LST and Raven's (Birney et al., 2012) finds that the LST has a strong correlation to Raven's, but in general, the LST appears to sit somewhere between Gf and WM tasks. It has the relational integration components shared between WM and Gf (as well as the additional, lower-order demands listed above) but it does not have the rule induction component necessary to extend the task to Gf. An important caveat to this finding that may concern some readers is that the LST and Raven's do, overall, correlate. However, the LST does not correlate more as the relational integration demands (as measured through RC) increase. Although this topic is explored more in our study on the LST (Chapter III), which indicates the RC demands in the LST may not entirely represent the demands of the task (despite being designed specifically as a manipulation of RC; Birney et al., 2006) for now, it is just worth pointing out that the increase in RC demands are not concomitant with the varying demands of Raven's. In Raven's, every problem only ever involves ternary relations, though difficulty can instead arise in the number of ternary relations that must be instantiated throughout a problem, or by the difficulty in identifying the ternary relations. Thus, it is not entirely unsurprising that increases in RC in the LST do not lead to increases in the correlation to Raven's (but again, this point is explored more in Chapter III).

One concern with the LST findings that is more valid is that the LST is a matrix task, so the LST and Raven's may overlap due to this surface similarity. However, the given-rule approach has also been successful outside matrix tasks. Oberauer et al. (2008) employed remarkably simple tasks such as the *finding squares* task and the *relation monitoring task* (this task is analysed and validated in Chapter V). In the finding squares task, participants are presented with a grid of blank dots which light up intermittently. The task is to respond whenever four dots in a grid all light up to form a square. In the monitoring task, participants are presented with a 3x3 grid of numbers which change intermittently. The task is to respond whenever a row or column in the grid matches some given match rule (e.g., all numbers in a row or column are even). In both tasks, the rule is provided. However, both tasks do require relational integration to bind the elements (the dots or the numbers) into a coordinated relation (a square or a set of even numbers). Consistent with the current theoretical perspective, Oberauer et al. (2008) found that these tasks (and similar ones with no rule induction) were the best predictors of Gf tasks, as compared to a large battery of more traditional WM tasks (such as complex span, which I have argued is indirectly measuring relational integration). Chuderski (2014) also utilized the relation monitoring task, finding it predicted Gf above-and-beyond other WM tasks. In addition, Chuderski found that increasing the number of bindings required in a relation decreased performance but did not influence the relationship to Gf (as with the LST).

In summary, it appears that rule induction is what makes a Gf task distinct from a WM task, with both based in relational integration. Given the apparent reduction in noise when rules are provided, it seems that rule induction simply complicates the measurement of WM capacity. The freedom to approach a problem such as Raven matrices with or without knowledge of rules means that control over participant performance is clouded. Similarly, to prevent the same rule being tested over and over, the participant must be able to keep track of attempted solutions and represent them as an independent chunk. These would initially occupy space in the direct access region until they have been committed to LTM, similar to the burden of a distracting task like subvocalizing, which may further cloud the relationship between WM and Gf.

Once we strip away rule induction, abstract reasoning tasks are essentially relational comparison tasks. These tasks typically manifest through rules such as those demonstrated in Figure 2.4. Despite the complex patterns and shapes that constitute a Raven’s problem (a ‘complex’ abstract reasoning task), the patterns and shapes boil down to relational rules such as these. Once the rule is known, the demand is in the binding of the elements into a series. Typically, this results in generating the element that continues the sequence, relying on other given information in the task (e.g., other shapes) or knowledge from LTM that dictates the range of a sequence (e.g., letters, numbers, a limited set of shapes).

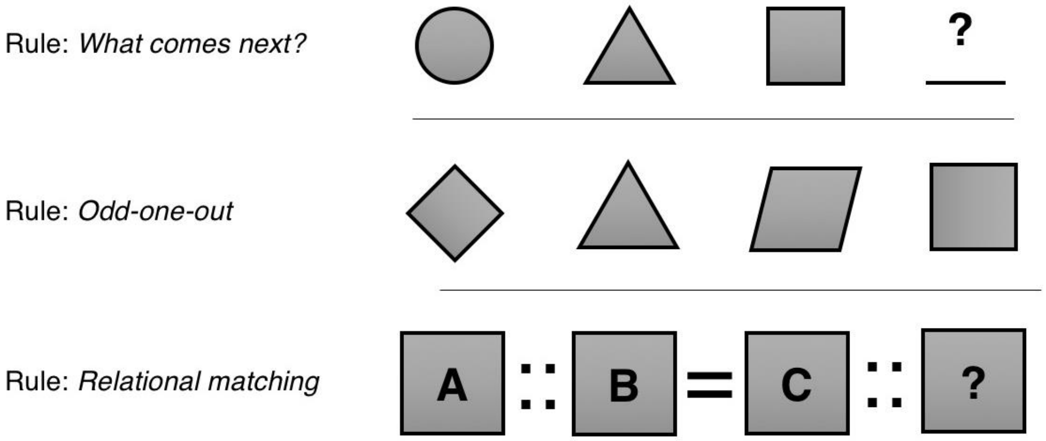


Figure 2.4. Example of common rules that require ‘abstract reasoning’. All abstract reasoning tasks employ a combination of one or more rules such as these that must be induced to solve each problem. In a relational integration theory of WM, the difficulty in these problems comes from having to bind all the elements into a coordination relation. Additional difficulty is provided by the rules being obscured or unknown at the outset, though this demand is distinct from the demands of relational integration and is what makes ‘abstract reasoning’ problems unique.

In summary, although higher-order Gf tasks are complex and multi-faceted, they can be fundamentally understood through demands on rule induction and relational integration, only the latter of which is shared by relational WM tasks.

III. STUDY 1: THE LATIN SQUARE TASK

The previous chapters argued that WM capacity can be best conceptualized in terms of relational integration demands (Halford et al., 1998) rather than just generic storage demands; and that WM and Gf are most fundamentally connected via the joint demands for relational integration. The dominance of storage-based WM tasks where quantity of recall is the primary outcome (Ackerman et al., 2005; Colom, Shih, et al., 2006) has (as argued here) obfuscated the interpretation of WM capacity. While WM is a system that stores information over time, this does not necessarily mean simple storage capacity (how many elements can be held at one time) should be the primary measure of WM capacity. For one, this assumes that WM tasks must involve storage over time (i.e., storing elements and recalling them up to several minutes later). Two, this does not account for the important role of chunking in measurement of capacity (Cowan, 2001). The field has been reluctant to shift beyond complex span tasks (Redick et al., 2012), in no part thanks to the rise of attentional control conceptualisations (Engle & Kane, 2004) where they are employed routinely. One problem with complex span tasks is that it is reasonably difficult to distinguish storage from attentional control demands within the task, since the performance outcomes (quantity of recall) tap both. Unsworth and Engle (2007a) have attempted to resolve the overlap in storage and attentional control demands within complex span tasks by distinguishing primary *active* storage from secondary *passive* storage. Active storage are elements within active attention while passive storage are those out of the focus of attention. Because active and passive storage are delineated by attentional activation, this distinction has been useful to attentional control theories. However, the complex span task does not benefit greatly from this distinction, because the nature of the task involves the frequent displacement of to-be-remembered elements from active to passive memory, making it difficult to determine the relative contribution of each type of storage to the end product (the recall). Although it could be argued that the frequent switching between active and passive characterizes the task demands, the different types of complex spans (e.g., reading span, operation span) have different processing tasks, some of which can be considered so simple (e.g., pattern judgements) that they may not actually draw the to-be-remembered elements out of active memory at all. Relational WM theories make a similar distinction between active and passive storage by considering the elements inside direct access (active) vs. those activated in long-term memory (passive), though it would primarily see the demands of the complex span as drawing from the capacity to form an ordering from the to-be-recalled elements (see Section 1.2). Thus, there is a need to consider tasks that more clearly distinguish active from passive storage, and that distinguish storage from relational processing. The current study illustrates once such attempt at disentangling these task components using the Latin Square Task (LST).

In Bateman (2015), I attempted to contrast active from passive storage demands in a variant of the LST by adding auxiliary storage demands to the task that were either active or passive in nature. The choice of task, despite its relevance to this chapter, was (at the time) somewhat tangential to this added storage-load manipulation, chosen mainly because it provided a way of contrasting the two storage loads in an auxiliary task (i.e., a storage load relevant to the task that must be kept active, and one irrelevant to the task that can be relegated to passive storage). In the LST (a more detailed description is provided in Section 3.1), participants are presented with a 4 x 4 matrix, partially filled with shapes (circle, square, etc.). Participants must deduce which shape should be in a marked target cell (signified by a ‘?’) using the one defining rule of the LST: *each row and each column may have only one of each shape*. This rule made it possible to operationalize the active and passive storage demands by contrasting whether the storage load (colour-marked cells of the matrix that had to be recalled after the actual item) was integral to the problem solution (active) or not (passive). Bateman (2015) found that if the added storage demands were passive, they were virtually irrelevant; both to main task performance and to relating the task to Gf (as measured through Raven's APM). The task became more difficult when performance was contingent on also recalling the added passive storage, but this still had no influence on the relationship to Gf. The added *active* storage demands meanwhile, both made the task more difficult and raised the correlation of the LST to Gf. In this condition, the added storage demands had to be kept active because they were involved in the main task. These results thus supported an attentional control explanation of WM-Gf, because it clearly demonstrated that added storage demands are only important to connecting the task to Gf if they are active in nature. Although this was not the original intention of the study, Bateman (2015) had somewhat ironically accomplished something using the LST (a relational integration task) that attentional control theories struggled with using the complex-span[6] (though see Dilevski, 2016, for a more direct response to this research question, manipulating the complex-span). There were however, two main problems with concluding Bateman's (2015) results using attentional control theories. First, is that the LST in itself is a relational integration task, not an attentional control task. This means that a more appropriate conclusion may be that the LST is relating to Gf because it measures relational integration, and the auxiliary attentional-control demands enhance this relationship, whereas auxiliary passive storage demands will not (also see Chapter V on the Relation Monitoring Task, which concludes a remarkably similar result). The second, far more pressing problem with an attentional control conclusion was that the added auxiliary demands were not the only manipulation of attentional control in this experiment. There was another manipulation of attentional control that *reduced* the attentional control demands of the task. The pressing problem with an attentional control conclusion is that the results of this other manipulation were in *direct contention* to this account. This other manipulation was

[6] It is also, of course, worth reminding the reader of the overlap between relational integration and attentional control theories frequently discussed in Chapter II.

‘Dynamic Completion’ (DC)[7], which reduced the attentional control demands involved in the task by allowing partial solutions to be offloaded onto the visual display, reducing the load associated with keeping shape elements in active memory. That is, participants could insert shapes into empty cells, effectively solving the puzzle in parts rather than only providing a single response to the target cell. Interestingly, in complete contrast to the result of the auxiliary demands, this DC version not just made the task considerably easier (as expected), but it *increased* the correlation of the LST to Gf. This result goes against the auxiliary demands manipulation, at least in (attentional control) theory. It also goes against the common wisdom that increases in task difficulty will always lead to concomitant increases in demand of Gf resources (Stankov, 2000; Stankov & Crawford, 1993). Given that this DC result is both surprising and interesting, the goal of the first study of this thesis was to replicate the DC finding and, if it could be successfully replicated, uncover more about how and why it manifests by comparing it to additional criterion measures (more than just Raven’s APM). We first consider the LST in more detail before moving on to the current experiments.

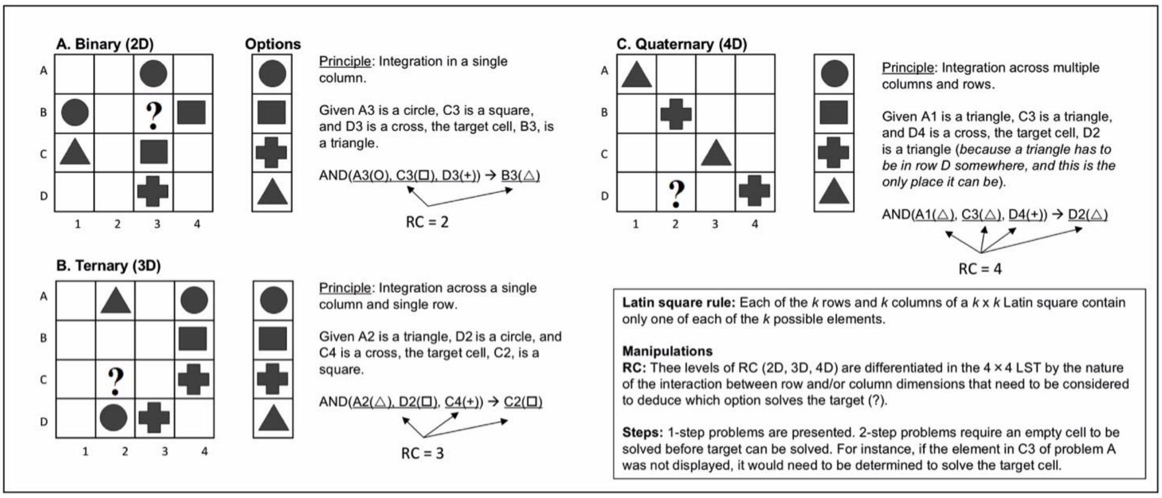
3.1. Introduction to the Latin Square Task

The LST was developed following the principles of RC theory (Birney, Halford, & Andrews, 2006; see also, Perret, Bailleux, & Dauvier, 2011; Zeuch, Holling, & Kuhn, 2011). Participants are presented with a partially filled 4 x 4 matrix (see Figure 3.1) that, when completed, contains exactly four instances of each of the four possible element types (typically circle, square, triangle, cross; but these could be a set of four colours, letters, or numbers, among other things) distributed according to the defining rule of a Latin square: that each row and each column must contain only *one* of each element type (because all experiments with the LST in this chapter use shapes, these elements are henceforth referred

[7] The term *dynamic completion* comes from Bowman (2006) who had earlier speculated on the benefits of employing such a manipulation to explore the LST.

to as shapes). The task is to deduce which shape fits a marked target cell according to this rule.

There are at least two distinct sources of cognitive demand in the LST. (1) Binding of to-be-integrated elements into a completed relation – the demand of which is indexed by RC; and (2) active storage demands associated with maintaining integrated, interim (sub-goal) outcomes for problems that involve multiple processing steps. These two demands are explicitly defined by item characteristics, as seen in the 'manipulations' inset of Figure 3.1. Put simply, the RC manipulation is defined by the number of row or column dimensions that must be integrated to come to the solution, notated in terms of dimensions (e.g., the two-dimensional binary problems are notated as '2D'); while the steps manipulation is the number of integration steps required to reach the target cell. If the two steps involved in a 2-step problem have different levels of RC (e.g., a 3D and a 4D), then the RC of that item is classified by the highest level of RC, in line with the axiom of RC theory that the overall complexity of a task is represented by the single most complex process involved in that task (Halford et al., 1998).



Principle: Integration in a single column.

Given A3 is a circle, C3 is a square, and D3 is a cross, the target cell, B3, is a triangle.

AND(A3(O), C3(□), D3(+)) → B3(△)

Principle: Integration across a single column and single row.

Given A2 is a triangle, D2 is a circle, and C4 is a cross, the target cell, C2, is a square.

AND(A2(△), D2(□), C4(+)) → C2(□)

Principle: Integration across multiple columns and rows.

Given A1 is a triangle, C3 is a triangle, and D4 is a cross, the target cell, D2 is a triangle (*because a triangle has to be in row D somewhere, and this is the only place it can be*).

AND(A1(△), C3(△), D4(+)) → D2(△)

Latin square rule: Each of the *k* rows and *k* columns of a *k* x *k* Latin square contain only one of each of the *k* possible elements.

Manipulations
RC: Thee levels of RC (2D, 3D, 4D) are differentiated in the 4 × 4 LST by the nature of the interaction between row and/or column dimensions that need to be considered to deduce which option solves the target (?).

Steps: 1-step problems are presented. 2-step problems require an empty cell to be solved before target can be solved. For instance, if the element in C3 of problem A was not displayed, it would need to be determined to solve the target cell.

Figure 3.1. Example LST items of three levels of complexity and associated RC analysis with underlining indicating independent dimensions to be integrated consistent with Birney, Halford, and Andrews (2006) using representational notation of Birney and Halford (2002).

Birney et al. (2006) found that the RC manipulation captures 64% of variability in item difficulty while the number of interim processing steps captures a further 16%, indicating that the task primarily loads on RC demands. Birney and Bowman (2009) investigated the RC and steps manipulations as a function of other tasks, such as Gf tasks as well as other RC and mental permutation tasks. Consistent with Birney et al. (2006), Birney and Bowman (2009) found evidence for a distinction between RC and interim steps. Higher Gf was associated with higher overall LST accuracy (average *r* = .47) however, contrary to expectations, this relationship was not moderated by RC level when collapsing over steps (η_p^2 = .01). That is, as RC increased, the relationship with Gf did not increase concomitantly, as would be expected by a complexity effect (Stankov, 2000). On the other hand, collapsing over RC, the relationship between Gf and LST performance was statistically greater for 2-step items than 1-step items (η_p^2 = .09). In other words, increasing the number of steps from one to two saw a sharp decline in performance for low Gf participants, but *not* for high Gf participants. Birney and Bowman concluded that the requirement for 'serial processing' was the component most likely linked to Gf. This conclusion aligns with theories where the overlap between WM and Gf represents attentional-control (Kane et al., 2004), in that interim information (the solution to the first step) must be kept active in the direct access region while processing on the next step is conducted. Because both steps must be solved to solve the problem, it is inextricable that a 2-step problem involves active storage of the interim solution (the outcome of the first step), as it must be used as part of the solution to the second step. In other words (and this following explanation will be relevant for thinking about DC), where a 1-step item involves integrating visually available shapes with the target cell, 2-step items involve integration of visually available shapes with the target cell *and* with a visually unavailable shape that is the outcome of the earlier processing step. This visually unavailable shape consumes capacity in the direct access region (aka active primary storage) because (a) losing it would result in restarting the problem; and (b) relegating it to LTM (aka passive secondary storage) would simply require reinstating it to the direct access region for use in the next step of the problem.

Thus, although the demands of the task seem more associated with RC (because the RC manipulation captures the vast majority of variability in item difficulty), the attentional control demands of the task (keeping visually unavailable representations active in direct access) seem more related to Gf. The future chapters will demonstrate that the lack of the RC by Gf covariation does not discredit relational integration but rather, contributes to the *relational integration hypothesis*: the theory that what fundamentally limits performance on Gf is relational integration, and even the most relationally simple version of the task can predict Gf just as well as more relationally complex versions – so long as they are still tapping relational integration. But even so, the significant Steps by Gf covariation (Birney & Bowman, 2009) seems to suggest that increased demands on attentional control can enhance the relationship to Gf. However, this interpretation is at odds with Bateman's (2015) DC findings, where *reducing* the attentional control demands involved in this active storage of the interim step *increased* the correlation to Gf. The current chapter sought to confirm the source of demands in the LST and how they relate to Gf by replicating the DC effect. If the DC effect does not replicate (i.e., DC does indeed reduce the correlation of the LST with Gf), then it would indicate that Bateman's (2015) finding was a one-off and attentional control is the critical component linking WM with Gf. On the other hand, if the associated Gf demands are in fact related solely to relational integration, we would expect the correlation between the LST and Gf to maintain in spite of the DC manipulation. This is because, where the standard LST entails relational integration demands (in generating the solutions to each step) and attentional control demands related to keeping the interim step outcomes active, the DC version removes these attentional control demands, leaving the only critical demand of the

DC version of the task as relational integration. An improvement in WM-Gf correlation would indicate that the LST has been purified as a measure of relational integration, lessening noise associated with attentional control demands (which, in this scenario, are actually detrimental to the relationship).

3.1.1. Aims and hypotheses

Because the DC findings appear to clash both with the 2-step findings of Birney and Bowman (2009) and the auxiliary cost findings in Bateman (2015), there is a clear need both of replication and of further theoretical analysis of the LST. The first study aimed to replicate the DC findings and add additional criterion WM measures to help determine the source of demands in the LST. In the first experiment, we included an additional WM measure (a complex-span) in addition to our Gf measure, Raven's APM. The complex span task chosen, the *symmetry span* (SSPAN), involves alternating between remembering an element and solving a basic processing judgement, then recalling the series of elements in order at the end of each trial. This quantifies WM capacity demands but, as discussed (in Section 1.2 and earlier in this chapter) may tap either active or passive storage (or some combination of the two). If the Basic version of the LST (i.e., the standard LST seen elsewhere) correlates more with SSPAN than the DC version, it would indicate that the variance being taken out of the DC version is indeed related to attentional control demands.

An additional issue with Bateman (2015) was the ordering of the LST blocks. Due to the layering of the instructions (with complex instructions for the auxiliary load items), the DC block always came after the Basic (standard LST) block. It is thus possible that the Basic block predicted less variance in Gf than DC because of the additional noise attributed to on-task learning induced by first exposure to the task. The DC block may have been a 'purer' measure simply because all participants had had experience with the task by that point. Learning has also been associated with APM performance (Bui & Birney, 2014; Lilienthal, Tamez, Myerson, & Hale, 2013) but it is nonetheless an experimental issue to have the blocks presented in a fixed order. As such, the present study would counterbalance the block order by presenting them in a random order.

Finally, the present study also recorded the number of ‘moves’ (empty cells filled) made by participants in the DC condition. If elaborate multi-step solutions were enabling strategic participants to solve DC, then it may indicate that strategy is related to Gf, and we would expect to see those scoring high in APM to be employing many moves in the DC condition.

Thus, the present study has three novelties over Bateman (2015): the inclusion of the SSPAN, the counterbalancing of block order, and recording of the number of moves. However, overall, the main purpose of the experiment is to simply replicate the remarkable findings of my earlier work. For the hypotheses, the standard descriptive effects of RC and Steps were expected: increases in RC and Steps would each result in linear decreases in performance. Consistent with Bateman (2015), it was hypothesized that the DC variant of the LST will increase task performance compared to the standard version, and that this increase will be more pronounced (an interaction) for 2-step items compared to 1-step items (since the benefit of LST-DC is theorized to primarily apply to 2-step items). Also consistent with Bateman (2015), it was hypothesized that the DC effect would replicate such that the correlation between the LST and Gf will increase when using LST-DC rather than LST-Basic, although both versions will correlate with Gf. The novel hypotheses were related to the additional classic WM measure added to the experiment, SSPAN. It was hypothesized that the difference in variability between the LST-Basic and LST-DC would be related to this classic WM measure, which is a task more representative of attentional control. In line with this, it was also hypothesized that the LST-DC would still correlate with Gf, above and beyond the variance already accounted for by the SSPAN. Further, in line with the relational integration hypothesis, we expect to see no relationship between the average number of DC moves and APM performance. If we do, it would indicate that the relationship with DC may form through strategy, rather than a purification of the relational integration demands.

3.2. Experiment 1: Method

3.2.1. Participants

The participants were 125 first-year psychology students at the University of Sydney who participated in exchange for course credit. There were 30 males and 95 females (76%) with an average age of 20.14 (SD = 4.43) years. Data from these participants are also reported in Chapter IV, though the focus of that chapter is the Arithmetic Chain Task (not reported here) rather than the LST.

3.2.2. Measures

Participants completed computerised versions of the LST (experimentally manipulated to include Basic and DC items), symmetry complex span (SSPAN), and Raven's Advanced Progressive Matrices (APM). Participants also completed the Arithmetic Chain Task though these results are not reported here. All tasks were programmed with *Inquisit Lab 4* (Millisecond Software, 2014).

Latin-Square Task (LST)

Participants were presented 24 items adapted from Birney and Bowman (2009) across two blocks (basic and DC). All items used the same four shapes as element types (circle, triangle, square, cross) and each had a 2-minute time-limit (with a countdown displayed to participants). If the time expired, the item was recorded as incorrect and the next item presented. Bateman (2015) reported that only 0.2% of all LST items attempted were marked incorrect through timeout, indicating that 2 minutes is sufficient. Each item had an RC (RC=2D/3D/4D) and steps (steps=1S/2S) combination (e.g., 2D-1S). Each block had 12 items with an equal distribution of RC*steps combinations (two of each combination).

Participants completed the basic block and the DC block in a random order, with the blocks visually differentiated by screen colour (white for basic; green for DC). There were separate instructions related to each (along with practice items) presented at the beginning of the experiment.

LST Basic. Basic items included the matrix and response options in the centre of the screen (see Figure 3.2). When a participant clicked a shape in the response options, the background of the selected shape would turn pink and *reset/confirm* buttons would appear. This gave participants a chance to confirm or change their response before moving on to the next item.

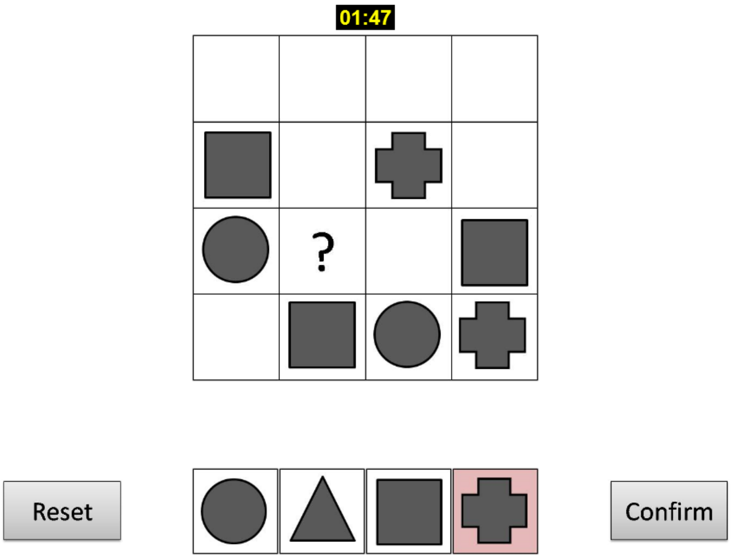
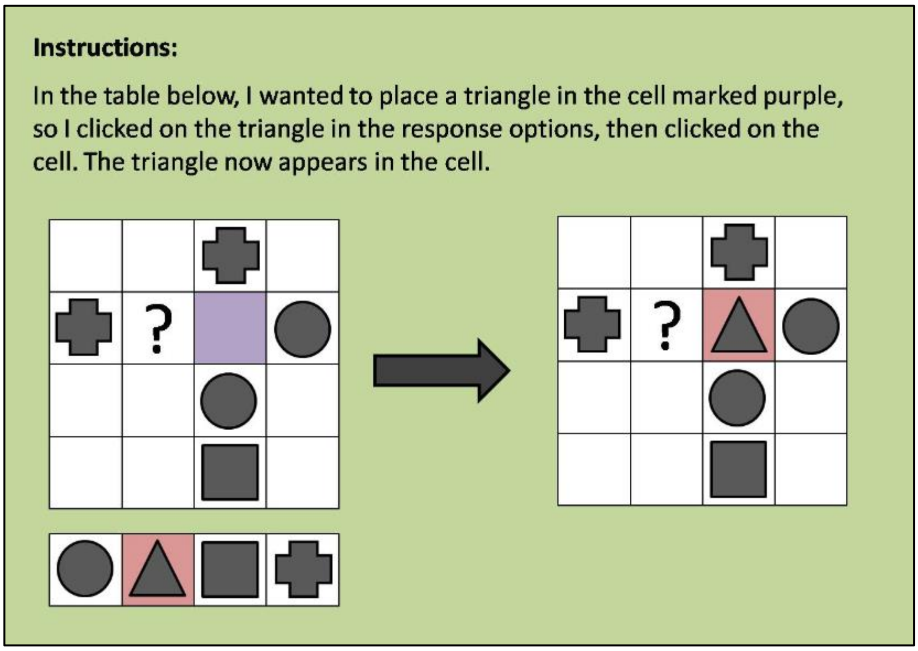


Figure 3.2. Example Basic item, as presented to participants.

LST Dynamic Completion (DC). DC items allowed participants to fill in the matrix before solving the target cell. To fill a cell in the matrix, participants selected their desired shape from the options then clicked on an empty cell (see Figure 3.3). The shape would appear in the cell and the cell background would change to pink to indicate it was an interim shape they had inserted. Participants could fill as many cells as they wished, though the instructions asked they only fill as many cells as necessary. A ‘move’ was recorded as any time a cell was filled by the participant and moves continued to cumulate regardless of resets.

Participants indicated their solution by placing the desired shape into the target cell in the same way as an empty cell (in this way, the minimum moves for each item was one). Only when a shape was placed into the target cell would the *reset/confirm* buttons appear for them to confirm their answer.



Instructions:

In the table below, I wanted to place a triangle in the cell marked purple, so I clicked on the triangle in the response options, then clicked on the cell. The triangle now appears in the cell.

Figure 3.3. Slide from the DC instructions, demonstrating an example item.

Symmetry Span

Participants were presented with alternating storage and processing tasks, as in Kane et al. (2004). For the storage task, participants viewed a 4x4 grid where a sequence of red squares would appear in one of the 16 potential locations. Two to five squares would appear in each set with each square appearing for 850ms. For the processing task, participants judged whether a displayed pattern was symmetrical along the vertical axis. Participants first viewed one square in the set, then completed a symmetry judgement, then viewed another square, then made another symmetry judgement, and so on until the entire storage set of squares had been displayed. After solving the last symmetry judgement for that set, participants would attempt to recall the squares in the order they were presented. The score analysed was the total number of correctly recalled squares across the task (2 x each set size), resulting in a possible range of 0-28. This partial scoring was favoured over a ‘span’ score (number of recalled squares within correctly recalled sets only) as it captures more variance (Redick et al., 2012).

Raven’s Advanced Progressive Matrices (APM)

Fluid intelligence was measured using a shortened 20-item version (odd items + items 34 and 36) of set II of the APM (J. C. Raven, 1941). Participants had 20 minutes to solve as many items as possible. This 20-item version has shown excellent reliability as a shortened version of the APM as it is sufficient for participants to learn and apply the rules that govern APM items (Bui & Birney, 2014).

Although a single task defines a construct such as fluid intelligence narrowly, this task was also chosen for its important surface and structural similarities to the LST. Both tasks employ a visuo-spatial matrix layout, and both are based on relational integration. The LST differs in that there is a single rule known to participants (the defining rule that only one of each type of shape can appear in each row and each column), while APM involves several unknown rules (Carpenter et al., 1990) that the participant must induce. In Chapter II, rule induction was identified as a defining characteristic of abstract reasoning tasks which set them apart from relational integration tasks such as the LST. However, APM elements are also generally more complex. Where the LST involves the same set of shapes (circle, triangle, square, cross) each time, APM elements are complex, with each element in a cell composed of multiple features. For instance, lines may, *inter alia*, be straight, wavy, dotted, and/or differ in orientation; shapes may, *inter alia*, differ in size, shading, numerosity, and/or form. Element complexity such as this is necessary to ensure the rules are being generalized across features. In the LST, changing the elements between items is unnecessary because the rules are given each time. Thus, although element complexity can be absolved into rule induction, it is important to consider that the difference in complexity between the two tasks could contribute to additional discrepancy in their intercorrelation.

3.3. Experiment 1: Results

3.3.1. Overview of the analyses

In line with the hypotheses, we first sought to determine the impact of RC (2D/3D/4D), Steps (1S/2S), and Condition (Basic/DC) on performance. To account for the low item size when breaking down the task by all three variables (which would be only two items per cell), the analyses are separated into one standard LST 'family' of effects (with the standard interaction of RC by Steps); then two DC 'families' that would be the focus of the DC hypotheses (RC by Condition; Steps by Condition). After the performance effects (accuracy), we then present the influence of Gf (through APM) on these performance effects. For replicating the standard LST family (RC and Steps), this is done using an ANCOVA because the RC and Steps variables are both theoretically continuous variables. For the Basic vs. DC comparison, this is done using a multiple linear regression because Basic and DC are not continuous. Instead, theoretically, Basic should constitute the same cognitive processes as DC, plus processes associated with additional attentional control demands. Thus, a regression is more appropriate for this comparison. We first begin with a presentation of descriptives.

3.3.2. Descriptives and correlations

Descriptive statistics for the tasks are provided in Table 3.1. Overall, the LST correlated with APM (r = .38, p < .001), and the DC condition had a slightly higher correlation to APM (r = .37, p < .001) than the Basic condition had to APM (r = .33, p < .001). As expected, the SSPAN correlated with LST-Basic (r = .29, p = .001) but not LST-DC (r = .14, p = .137). The SSPAN also correlated with APM (r = .25, p = .005).

Table 3.1. *Descriptives (proportion correct) and correlation coefficients for task measures in Experiment 1.*

	Descriptives		Correlations			
	Mean	SD	DC	Total	*SSPAN*	APM
LST-Basic	0.84	0.15	**.62**	**.85**	**.29**	**.33**
LST-DC	0.89	0.15	**-**	**.90**	.14	**.37**
LST-Total	0.86	0.14		-	**.24**	**.38**
SSPAN	0.75	0.16			-	**.25**
APM	0.60	0.20				-

N=125; bold coefficients p < .05.

For moves, a sum total was calculated from all DC items. The average total number of moves was 48.42 (equating to an average of approximately four moves per item) but varied greatly between participants (SD = 29.72). This sum total moves was indeed positively correlated with performance in the DC condition (*r* = .30, p = .001), such that more moves led to, on average, higher performance; but the number of moves was not correlated with APM performance (*r* = .04, p = .124).

3.3.3. Performance effects

Consistent with Bateman (2015), there was a linear trend for RC, such that increasing complexity led to decreased performance ($F_{1,124}$ = 127.72, *mse* = 0.633, p < .001, η_p^2 = .507). The effect of Steps was also significant, such that 2-step items were more difficult than 1-step items ($F_{1,124}$ = 27.52, *mse* = 0.419, p < .001, η_p^2 = .182). The difference between Basic and DC on performance was also significant, ($F_{1,123}$ = 20.46, *mse* = 0.400, p < .001, η_p^2= .143) but the effect was not moderated by an interaction with the linear trend of RC ($F_{1,123}$ = 3.05, *mse* = 0.381, p = .083, η_p^2 = .024). In other words, the benefit of DC relative to Basic was not dependent on certain levels of RC. However, consistent with the hypotheses, there was a significant interaction between Condition and Steps ($F_{1,123}$ = 5.61, *mse* = 0.575, p = .019, η_p^2 = .044), such that 2-step items benefitted more from DC than 1-step items did. This interaction is depicted in Figure 3.4.

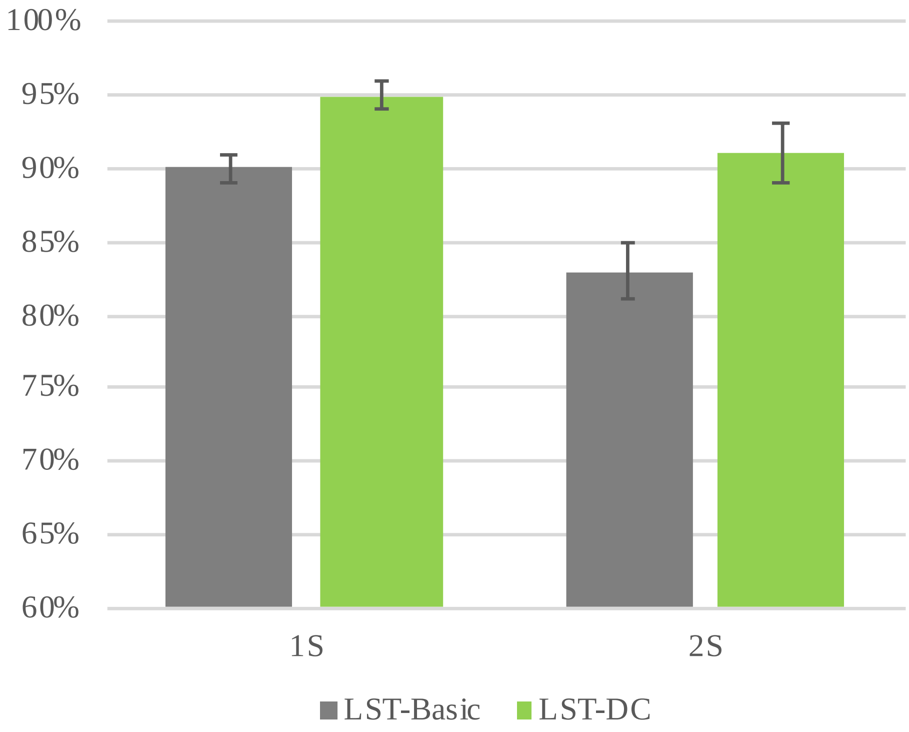


Figure 3.4. Mean proportion scores for LST conditions separated by 1-step and 2-step problems. Error bars indicate standard error (n=125).

To determine the impact of RC and Steps on the shared demands with Gf, APM was entered as a covariate into an ANCOVA with RC and Steps as repeated measures. This replicates prior research on the standard LST (RC by Steps) using the ANCOVA method (Birney & Bowman, 2009). The linear effect of Steps was moderated by APM ($F_{1,123}$ = 6.157, *mse* = 0.402, p = .014, η_p^2 = .048). Unlike past research (e.g., Birney & Bowman, 2009), the linear effect of RC was also moderated by APM, ($F_{1,123}$ = 7.480, *mse* = 0.604, p = .007, η_p^2 = .057), such that increasing RC significantly increased the covariation of the task to Gf. To determine the relative contribution of each RC level to this linear effect, an additional regression was run with each RC level entered separately into a model predicting APM. From this regression, it was clear that the linear effect was being carried by 2D and 4D items. That is, the 2D items significantly predicted APM (ΔR^2 = .086, p = .001) but the 3D items contributed virtually nothing additional (ΔR^2 < .001, p = .986). The 4D items were then a significant predictor above the other two levels (ΔR^2 = .086, p < .001). In the final model, only 4D items were significant predictors of APM.

3.3.4. Basic and DC regressions

To determine the relative, unique contribution of Basic and DC against each other, the two conditions were entered sequentially (Basic first, then DC) into a regression predicting APM performance. The first step, with just Basic, accounted for 10.8% of variance in APM (R^2 = .108, p < .001). Adding DC in the second step was a significant change (ΔR^2 = .046, p = .012) and reduced the unique contribution of Basic to non-significance in this final model (Basic sr^2 = .02, p = .137; DC sr^2 = .05, p = .012).

Next, the regression predicting APM was repeated except WM was controlled for by adding in SSPAN as an initial step (before LST-Basic). The results of this regression are provided in Table 3.2. SSPAN was added as the first step and this SSPAN-only model was significant (R^2 = .063, p = .005). Adding Basic in a second step was a significant change (ΔR^2 = .072, p = .002) and this addition caused the SSPAN to no longer have a significant unique contribution (SSPAN sr^2 = .03, p = .060; Basic sr^2 = .07, p = .002). This indicates that the variance associated with SSPAN (theorized to be storage-related) was subsumed by Basic, which itself added significant meaningful variance (theorized to be processing-related). Finally, adding DC in a third step was also a significant change (ΔR^2 = .050, p = .008); and one that subsumed the unique contribution of Basic. Interestingly, this addition caused the SSPAN to once more have a significant unique contribution (SSPAN sr^2 = .03, p = .037; Basic sr^2 < .01, p = .362; DC sr^2 = .05, p = .008). Thus, the attentional control component shared by SSPAN and Basic was prioritized in the SSPAN; while the shared LST (relational) components in Basic and DC was prioritized by DC.

Table 3.2. *Regression Model Predicting Gf with DC in Experiment 1.*

Model	Predictor	B	*t*	*p*	*sr*^{2}	R^2	ΔR^2
1	Symmetry Span	**.221**	2.84	.005	.063	.063	**.063**
2	Symmetry Span	**.149**	1.90	.060	.026	.135	**.072**
	LST-Basic	**.633**	3.16	<.002	.072		
3	Symmetry Span	**.161**	2.11	.037	.031	.185	**.050**
	LST-Basic	**.226**	0.91	.362	.006		
	LST-DC	**.644**	2.70	.008	.050		

N=125; bold coefficients p < .05.

3.4. Experiment 1: Discussion

Overall, the results of Experiment 1 replicated the DC purification effect of Bateman (2015). That is, although LST performance increased as a result of the DC manipulation, the relationship between the LST and APM also increased significantly. We also observed an expected Steps interaction with DC, such that 2-step items benefitted from DC more than 1-step items. Although this is expected from DC, this interaction was not seen in Bateman (2015). Nonetheless, both experiments demonstrated a clear purification effect: the LST predicts APM better when attentional control demands are minimized by way of DC. These attentional-storage demands were further supported conceptually by a correlation between LST-Basic and SSPAN (a typical storage-focused WM task) that was not also seen between LST-DC and SSPAN. Indeed, the results of the first and second models of the regression predicting APM indicated shared variance between LST-Basic and SSPAN. Thus, LST-Basic does not appear to simply be a worse measure than LST-DC, it simply assess additional processes that overlap with complex spans, which we have observed may contribute noise to the prediction of Gf. The LST-DC appeared to tap a unique demand related to APM which we have theorized is a pure measure of relational integration.

One possibility that we explored was that LST-DC was allowing for new strategies to emerge, as there are now more feasible ways to solve each problem which are otherwise hard-limited by the intense storage demands of multi-step pathways. We investigated this possibility by analysing the number of moves made by participants. Although more moves generally led to better performance (as would be expected by additional solution pathways), the number of moves was not related to APM. Thus, the relationship between LST-DC and APM does not simply seem to be about enabling new strategies.

A final point worth considering is that the linear RC effect did actually covary with Gf. That is, as RC rose, so did the correlation with APM. This was unexpected given prior research (Birney & Bowman, 2009) has failed to find this covariance, and because the RC analysis of the APM determined that the APM does not vary by RC. Given these inconsistent results, further research is needed to confirm if this finding can be replicated.

Although the remarkable DC finding was replicated, the experiment was still somewhat limited by the number of tasks involved. The APM is a well-accepted measure of Gf (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002) and the task format overlap (matrix-style) helped to isolate the DC effect from the Basic version (because all three tasks involve matrices). Nonetheless, the overlap in task format may concern some readers. For the next experiment, we added an additional verbal (i.e., non-matrix) Gf task, the Letter Series. We also added an additional WM measure, the *n*-back, and replaced the SSPAN with the Operation Span (OSPAN). The processing aspect of the SSPAN (judging the symmetry of patterns) can be completed using lower-order visual strategies, which is less processing-intensive than the OSPAN (judging the veracity of arithmetic operations). Thus, the SSPAN may have simpler attentional control demands because the to-be-remembered elements can be kept exclusively in active storage. The OSPAN may better tap an attentional control demand because it requires more intensive processing and thus, should require more attentional demands in the frequent shifts of to-be-remembered elements between active and passive storage. The OSPAN is also the more common complex-span (Redick & Lindsey, 2013) and tends to have higher reliability than the Symmetry Span (Redick et al., 2012).

Finally, because our main theory for LST-DC comes down to a ‘purification’ effect, we also considered an established relational integration task, the Relation Monitoring Task (RMT) (Oberauer et al., 2008). While the LST was designed and validated by RC theory (Birney et al., 2006), the DC variant is nonetheless still unknown. The RMT meanwhile, has shown remarkable success as a relational integration measure (Oberauer et al., 2008, Chuderski, 2014). Chapter V details the RMT in-depth but for now, the RMT is a suitable task as a pure measure of relational integration. We should expect that if the LST-DC is indeed drawing out the relational integration capabilities of the LST, then it should correlate with the RMT, which should in turn correlate with Gf.

Once again, we predicted that DC would lead to an increase in performance and an increase in the correlation to Gf, as measured through APM and Letter Series, in comparison to Basic. We also expected the LST-DC to correlate with the RMT more than LST-Basic, but the LST-Basic would correlate more with the WM measures (*n*-back and OSPAN) than the LST-DC would.

3.5. Experiment 2: Method

3.5.1. Participants

One-hundred participants (67 female, 33 male) took part in exchange for course credit. Their average age was 19.47 (SD = 2.12) years. Participants undertook six tasks: the LST (with Basic and DC blocks), two measures of WM (Operation Span, spatial *n*-back), two measures of Gf (APM, Letter Series), as well as the RMT, a measure of relational integration. Participants completed the tasks in a random order in 90-minute sessions, in groups of up to eight in computer labs at the University of Sydney. Data from these participants are also reported in Chapter V, though the focus of that chapter is experimental manipulations of the Relation Monitoring Task rather than the LST. In this experiment, we only consider total RMT scores.

3.5.2. Measures

The same LST and APM from Experiment 1 was employed. The OSPAN replaced the SSPAN. We also added three additional criterion measures: the *spatial n-back* for traditional WM, the *Letter Series* for Gf, and the *Relation Monitoring Task* for relational WM. These additional tasks are described below.

Operation Span

Participants completed the Operation Span (OSPAN) with set sizes of 3, 4, 5, and 6 (two sets of each). In each set, participants alternated between memorizing a letter and verifying the truth of a mathematical operation. Once all letters for that set had been presented, participants attempted to recall the letters in the order they were presented. Again, partial credit scoring using the total number of correct letters (*OSPAN Letters*) was preferred to capture more variance (Redick et al., 2012).

Spatial n-back

Participants viewed a 3x3 grid where a sequence of blue squares would appear in one of nine potential locations. The participants' task was to respond when a square appeared in the same location as the trial *n* back from the current location. Each square appeared for 500ms and there was a 2500ms interlude between each square, resulting in trial durations of three seconds. Participants received two 2-back blocks and two 3-back blocks. Each block consisted of 14 non-target trials and six target trials. The score analysed was the ratio of percentage of hits to match trials divided by percentage of false alarms to no-match trials, averaged across blocks.

Letter Series

Participants had four minutes to complete as many of 15 Letter Series items as they could. Each item involved a patterned sequence of letters followed by an underscore to indicate that the task was to complete the pattern by inserting a single letter to the end of the sequence. Like APM, the items become progressively more difficult.

Relation Monitoring Task

The RMT (Oberauer et al., 2008) involved presenting a continuous 3 x 3 array of 3-digit number strings. The task was to respond (with the spacebar) whenever an array matching the current match rule was presented. If the array did not match the current rule, the participant was to wait for the next array, which would replace some or all strings (depending on the condition) with new ones. Each array was presented for 5.5 seconds with a 100ms interval. Although there were several experimental manipulations administered in the RMT, these are not relevant here and as such, only the aggregate score is considered. For details on these manipulations, see Chapter V.

3.6. Experiment 2: Results

3.6.1. Overview of the analyses

As was the case for Experiment 1, we begin the analyses by presenting descriptives and correlations, then performance effects of the LST manipulations. We then consider the influence of Gf on these performance effects using ANCOVAs (and an additional regression on the RC levels to consider non-continuous influence on RC). Finally, a series of regressions are conducted comparing the relative contribution of Basic and DC on predicting *Gf* while controlling for the criterion WM measures (OSPAN, *n*-back, RMT).

3.6.2. Descriptives and correlations

Descriptive statistics for Experiment 2 are provided in Table 3.3. Once again, the LST overall correlated with APM and it was stronger than in the prior experiment (r = .54, p < .001). As seen in Table 3.3, both Basic and DC had strong correlations with APM, though this time Basic had the higher correlation with APM (r = .51, p < .001), as compared to DC with APM (r = .46, p < .001). A similar pattern was seen in correlating the LST to Letter

Series: Basic (*r* = .46, p < .001) and DC (*r* = .40, p < .001). The two Gf measures, APM and Letter Series, correlated with each other (*r* = .42, p < .001), and together formed a *Gf* variable using principal axis factoring with varimax rotation (this variable accounted for 70.88% of variance in the two measures). The OSPAN and *n*-back did not correlate with each other (*r* = -.02, p = .817), but each correlated separately with *Gf* (OSPAN~Gf *r* = .23, p = .023; *n*-back~Gf *r* = .49, p < .001). The OSPAN also did not correlate with either LST condition (Basic *r* = .05; DC *r* = .06) but did correlate with APM to a similar degree as the SSPAN in Experiment 1 (*r* = .24, p = .017); while the *n*-back correlated with both LST conditions (Basic *r* =.44; DC *r* = .41). Finally, the RMT correlated with both LST measures (Basic *r* =.46; DC *r* = .45) and correlated highly with *Gf* (*r* = .59).

Table 3.3. *Descriptives (proportion correct) and correlation coefficients for task measures in Experiment 2.*

	Descriptives		Correlations						
	Mean	SD	DC	OSPAN	*n*-back	L-Series	APM	*Gf*	RMT
Basic	.80	.18	**.63**	.05	**.44**	**.45**	**.51**	**.57**	**.46**
DC	.83	.18	**-**	.06	**.41**	**.40**	**.46**	**.51**	**.45**
OSPAN	.83	.18		**-**	-.02	.15	**.24**	**.23**	.13
n-back	2.42*	1.64*			-	**.51**	**.32**	**.49**	**.44**
L-Series	.69	.12				**-**	**.42**	**.84**	**.53**
APM	.61	.19					-	**.84**	**.47**
Gf	0.00*	0.77*						-	**.59**
RMT	.67	.13							-

N=100; bold coefficients p < .05.

*n-back mean and SD based on block-average hits minus false alarms, rather than proportion correct; *Gf* mean and SD based on factor score, rather than proportion correct.

3.6.3. Performance effects

Consistent with prior experiments, there was a linear trend for RC, such that increasing complexity led to decreases in performance ($F_{1,87}$ = 103.86, *mse* = 0.819, p < .001, η_p^2 = .544). There was also a linear trend for Steps, such that 2-step items were more difficult than 1-step items ($F_{1,87}$ = 23.86, *mse* = 0.355, p < .001, η_p^2 = .215). For the novel family effects, DC resulted in significantly higher performance than Basic ($F_{1,87}$ = 9.10, *mse* = 0.460, p = .003, η_p^2 = .095).

Unfortunately, due to a programming error in this experiment, the item-level data of the LST was not recorded. That is, only composite scores of each RC level, averaged over all levels of Steps (and vice-versa), were recorded, rather than specific RC by specific Steps composites. In other words, the data provided composites such as “4D items” rather than composites such as “4D1S items”. This meant the interactions and covariation with Gf could not be perfectly replicated from the prior experiment (where RC and Steps were entered together in a single ANCOVA with Gf). Rather, each variable could only be entered separately, averaged across the other variable. This meant that, in general, the effects would be overestimated as compared to Experiment 1, since the measures had a larger range (because each measure constituted more items as a result of being averaged over the non-considered manipulation). With this in mind, Steps significantly covaried with *Gf*, ($F_{1,96}$ = 12.722, *mse* = 1.369, p < .001, η_p^2 = .117). On its own, the linear effect of RC also significantly covaried with the *Gf* factor, ($F_{1,96}$ = 52.662, *mse* = 1.073, p < .001, η_p^2 = .354), such that increasing RC led to higher correlations with Gf. In a regression predicting *Gf* using the separate RC levels, each subsequent RC level both increased R^2 significantly and subsumed the contribution of prior RC levels. In the final model, only 4D items were contributing uniquely (Overall R^2 = .410; 2D sr^2 < .01; 3D sr^2 < .01; 4D sr^2 = .20).

3.6.4. Basic and DC regressions

For condition (Basic vs. DC), the same regression approach as with the prior experiment was used, except this time the predictor tasks were predicting the *Gf* factor rather than APM alone. The pattern of regression results using LST alone was similar to prior experiments, with LST-Basic accounting for 32% of variance in *Gf* (R^2 = .321, p < .001), and adding DC significantly increased this to R^2 = .359 (ΔR^2 = 0.038, p = .020). In the final model, both Basic and DC provided unique contributions (Basic sr^2 = .10, p < .001; DC sr^2 = .04, p = .020), a somewhat different outcome to Experiment 1, where DC subsumed the unique contribution of Basic.

The next regression controlled for WM by adding OSPAN and *n*-back to a preliminary model which on its own did predict *Gf* (R^2 = .304, p < .001), with both measures providing unique contributions (OSPAN sr^2 = .06, p = .007; *n*-back sr^2 = .25, p < .001). Adding LST-Basic was a significant increase (ΔR^2 = .143, p < .001) though adding LST-DC on top of this was a marginally non-significant increase (ΔR^2 = .021, p = .061). Running the same model again without LST-Basic demonstrated that LST-Basic and LST-DC were largely contributing the same variance, as LST-DC became a significant change on its own, above the two WM measures (ΔR^2 = .108, p < .001). Finally, to determine the relative contribution of the RMT as a ‘pure’ relational integration measure, we added RMT to the predictors of the full model (OSPAN, *n*-back, Basic, and DC predicting *Gf*) to determine its relative impact on the existing predictors. The RMT was a significant increase in the variance predicting *Gf* (R^2 = .523, ΔR^2 = .056, p = .002), though did not change the significance of the unique contributions of the other predictors. The full results of this complete regression with the unique contributions of each predictor at each stage of the model is presented in Table 3.4. This Table also demonstrates that, unlike Experiment 1, the LST-Basic did not subsume the variance of the WM measures in Model 2.

Table 3.4. *Full Regression Model Predicting Gf in Experiment 2*

Model	Predictor	B	*t*	*p*	*sr*^{2}	R^2	ΔR^2
1	Operation Span	**.239**	2.76	.007	.057	.304	**.304**
	n-back	**.502**	5.80	< .001	.252		
2	Operation Span	**.206**	2.65	.010	.042	.446	**.143**
	n-back	**.316**	3.66	< .001	.081		
	LST-Basic	**.422**	4.87	< .001	.143		
3	Operation Span	**.200**	2.56	.011	.040	.467	.021
	n-back	**.286**	3.29	.001	.064		
	LST-Basic	**.316**	3.09	.003	.056		
	LST-DC	.190	1.90	.061	.021		
4	Operation Span	**.170**	2.30	.024	.028	.523	**.056**
	n-back	**.213**	2.49	.015	.033		
	LST-Basic	**.248**	2.50	.014	.033		
	LST-DC	.127	1.30	.196	.009		
	RMT	**.291**	3.24	.002	.056		

N=100; bold coefficients p < .05.

3.7. Experiment 2: Discussion

Overall, in Experiment 2, we still found encouraging results for the use of LST-DC, though they were slightly less remarkable than in Experiment 1. This time, the LST-Basic maintained a stronger position as a predictor of *Gf*, with the LST-DC adding a marginally non-significant contribution. The slightly lower sample size in this experiment (*n* = 100, compared to 125 in Experiments 1) may have resulted in a Type II error occurring on the LST-DC, though this in itself does not explain the now more substantial contribution of LST-Basic. In general, the LST-DC also correlated better with the WM measures (*n*-back and OSPAN) than it did in the earlier experiments. Despite these differences, the result is still largely in line with the overall core finding of Experiment 1: that a DC variant of the LST, which substantially reduces the difficulty by the task by minimizing demands on active storage, still correlates well with complex *Gf* tasks. In fact, the only reason the results of Experiment 2 are perhaps surprising is due to the remarkable results of Experiments 1 and

Bateman (2015). Taken in isolation, Experiment 2 still demonstrates the power of LST-DC, a version of the task with minimal attentional control demands.

Although the results of Experiment 2 are not completely divergent to the earlier experiment, the addition and replacement of several tasks nonetheless introduced some substantial changes to the experiment structure which may have affected the overall interpretability of the results, when taken together. For instance, although the SSPAN was criticized in Section 3.4 for having overly simple processing, the very reason it may have worked well as a correlate with LST-Basic and not LST-DC may be because of this simple processing. This simplified processing allows for more pure capture of active storage, rather than a switching demand associated with movement between active and passive storage which may be induced by the higher processing demands of arithmetic in the OSPAN. As such, the third experiment aimed to rectify doubts associated with the inconsistent task selection by including both the OSPAN and SSPAN, in addition to the *n*-back. Once again, both the APM and Letter Series acted as measures of Gf.

It was hypothesized that LST-DC would lead to an increase in performance and an increase in the correlation to Gf, as measured through APM and Letter Series, as compared to LST-Basic. It was also hypothesized that LST-Basic would correlate more with the WM measures (*n*-back, OSPAN, and SSPAN) than LST-DC would.

3.8. Experiment 3: Method

3.8.1. Participants

In total, 106 participants (74 females, 32 males) took part in exchange for course credit. The average age was 19.90 (SD = 3.85) years. Participants undertook seven tasks: the LST (with Basic and DC blocks), three measures of WM (Operation Span, Symmetry Span, and spatial *n*-back), two measures of Gf (APM, Letter Series), as well as the Swaps task, which is not reported here. Participants completed the tasks in a random order in 90-minute sessions, in groups of up to ten in computer labs at the University of Sydney. Data from these participants are also reported in Chapter VI, though the focus of that chapter is the experimental manipulations of the Swaps task rather than the LST.

3.8.2. Measures

The same LST, APM, Letter Series, OSPAN, SSPAN, and spatial *n*-back were employed as in prior experiments. Participants also completed the Swaps task (Stankov & Crawford, 1993), which is not reported here. Participants completed as many tasks as they could in the time provided. The tasks were presented in a random order, except for the SSPAN, which was always presented last. This was because the SSPAN was deemed the least necessary task. In total, 26 of the 106 participants did not reach the SSPAN. A small minority of these 26 participants also failed to complete at least one other task (four for LST, one for *n*-back, and one for both APM and OSPAN). The implications on the missing data from the SSPAN are described in the Results (Section 3.9.2).

3.9. Experiment 3: Results

3.9.1. Overview of the analyses

As was the case for the prior experiments, we begin the analyses by presenting descriptives and correlates, then performance effects of the LST manipulations. The influence of Gf on these performance effects are considered with ANCOVAs, then regressions are used to consider the relative contribution of Basic and DC on Gf.

3.9.2. Descriptives and correlations

As seen in Table 3.5, the means and standard deviations were generally as expected across the tasks. One exception of note is the difference between OSPAN and SSPAN: participants found the SSPAN considerably more difficult than the OSPAN, with an average of 3 less elements recalled in total. Because the same set sizes were used, there is no reason to suspect that SSPAN should be substantially more difficult. The more likely culprit was that because participants always completed the SSPAN last, they may have felt more fatigue or felt more rushed to complete the task to end the session on time. Although the SSPAN descriptives themselves were not overly concerning, the potential that the participants who reached the SSPAN represented a different subset of participants than those who did not reach the SSPAN was concerning. In general, the scores of those who reached the SSPAN were slightly higher than those who did not across the other tasks (e.g., Swaps M = .63 to .58). Although none of these differences reached significance in independent samples *t*-tests, this was largely due to the high variance from the small sample that did not reach the SSPAN (*n* = 26), rather than due to the negligibility of the mean differences. Thus, due to concerns over selection bias in sampling only those who reached the SSPAN, the easier course of action was to simply exclude the SSPAN from the analyses to ensure the full sample of 106 was used whenever possible. This was only possible because the SSPAN was always presented last and thus, did not contaminate the earlier tasks. Where possible, data on the SSPAN is still mentioned, but their findings should be interpreted cautiously.

Once again, the LST overall correlated with APM (*r* = .45, p < .001). As seen in Table 3.5, both Basic and DC had strong correlations with APM, with DC having a slightly higher correlation with APM (*r* = .42, p < .001) than Basic with APM (*r* = .39, p < .001). This time, the Letter Series was more weakly, but still significantly, correlated with Basic (*r* = .21, p = .036) and DC (*r* = .23, p = .020). However, the two Gf measures, APM and Letter Series, still correlated with each to a remarkably similar extent as the prior experiment, (*r* = .42, p < .001). Together, a *Gf* variable was formed using principal axis factoring with varimax rotation, accounting for 71.09% of variance in these two Gf measures. The OSPAN and SSPAN were weakly correlated with each other (*r* = .22, p = .048). Each were more strongly correlated with the *n*-back, with the SSPAN and *n*-back correlation (*r* = .46, p < .001) being somewhat higher than the OSPAN and *n*-back (*r* = .32, p = .001), potentially attributable to the spatial nature of the tasks but potentially also attributable to the subsample (n = 80) of those that completed all tasks. The SSPAN had a strong correlation to *Gf* (r = .55, p < .001), while the OSPAN had a weaker but significant correlation to *Gf* (r = .20, p = .044). In general, most correlations with *Gf* improved slightly as a result of reducing the sample to only the 80 who completed all tasks, despite the reduction in sample power.

Table 3.5. *Descriptives (proportion correct) and correlation coefficients for task measures in Experiment 3.*

	Descriptives		Correlations						
	Mean	SD	DC	OSPAN	SSPAN	*n*-back	LSeries	APM	*Gf*
Basic	.81	.17	**.67**	.18	**.25**	**.43**	**.21**	**.39**	**.37**
DC	.88	.16	**-**	.08	**.26**	**.44**	**.23**	**.41**	**.38**
OSPAN	.90	.14		**-**	**.22**	**.32**	.19	.15	**.20**
SSPAN**	.78	.17			-	**.46**	**.32**	**.57**	**.55**
n-back	2.63*	1.68*				-	**.40**	**.48**	**.52**
L-Series	.68	.15					**-**	**.42**	**.84**
APM	.63	.18						-	**.84**
Gf	0.00*	0.77*							-

N=106; **SSPAN sample N = 80; bold coefficients p < .05.
*n-back mean and SD based on block-average hits minus false alarms, rather than proportion correct; *Gf* mean and SD based on factor score, rather than proportion correct.

3.9.3. Performance effects

Once more, we begin the analyses with a replication family (RC by Steps) followed by two novel families testing the Basic vs. DC comparison (RC by Condition; Steps by Condition). Consistent with prior experiments, there was a linear trend for RC, such that increasing complexity led to decreases in performance ($F_{1,100}$ = 119.71, *mse* = 0.591, p < .001, partial-η^2 = .545). There was also a linear trend for Steps, such that 2-step items were more difficult than 1-step items ($F_{1,100}$ = 35.20, *mse* = 0.414, p < .001, partial-η^2 = .260). For the novel family effects, DC resulted in significantly higher performance than Basic ($F_{1,100}$ = 21.60, *mse* = 0.611, p < .001, partial-η^2 = .178). Contrary to Experiment 1, there was a (marginally) significant interaction between this linear trend for RC and Condition, such that the linear trend of RC was more pronounced for Basic than for DC ($F_{1,100}$ = 4.05, *mse* = 0.413, p = .047, partial-η^2 = .039). Also contrary to Experiment 1, there was no interaction of condition with the linear trend of Steps ($F_{1,100}$ = 0.33, *mse* = 0.610, p = .568).

To investigate covariation, the same approach as the prior experiments was used, with ANCOVAs for the RC and Steps effects. Steps covaried with *Gf* ($F_{1,97}$ = 5.422, *mse* = 0.395, p = .022, η_p^2 = .053), with increased correlation to Gf for 2-step items compared to 1-step items. The linear effect of RC also again covaried with *Gf* ($F_{1,97}$ = 7.966, *mse* = 0.549, p = .006, η_p^2 = .076), with higher RC levels having a higher correlation to Gf. However, once again, the regression isolating the effect of each RC level replicated Experiment 1 with 4D items producing a unique contribution to the model (sr^2 = .06), while 2D and 3D items appeared to be sharing variance because 3D items did not significantly improve the model (ΔR^2 = .020, p = .139) that contained 2D items alone (R^2 = .092, *p* = .002).

3.9.4. Basic and DC regressions

For the regression with conditions, this time, the LST-Basic alone accounted for a smaller (relative to prior experiments) but still significant 14% of variance in *Gf* (R^2 = .135, p < .001), and adding DC was a marginally non-significant increase to R^2 = .166 (ΔR^2 = 0.031, p = .060). In the final model, neither Basic nor DC provided unique contributions (Basic sr^2 = .02, p = .133; DC sr^2 = .03, p = .060). Although this was different to the prior experiment, it was clear that these results may have been due to the now considerably weaker correlation between Letter Series and the LST. The correlation between Letter Series and APM gave no reason to be suspicious of the Letter Series acting strangely in this experiment, but the regression was nonetheless repeated using just APM as the dependent variable, to help situate the results in the context of all experiments. With this approach, the LST-Basic alone accounted for a significant 15% of variance in APM (R^2 = .149, p < .001). Adding the DC increased this significantly to 20% (ΔR^2 = 0.055, p = .012). In the final model, only DC uniquely contributed significantly (Basic $sr^2 = .02$, $p = .092$; DC $sr^2 = .05$, $p = .012$). Thus, the regression replicated prior experiments, but only when considering APM in isolation as the Gf indicator.

Next, we controlled for WM by adding OSPAN and *n*-back to a preliminary model, which on its own predicted the *Gf* factor ($R^2 = .339$, $p < .001$), with *n*-back providing a unique contribution, but not OSPAN (OSPAN $sr^2 < .01$, $p = .423$; *n*-back $sr^2 = .27$, $p < .001$). Adding LST-Basic was a marginally non-significant increase ($\Delta R^2 = .024$, $p = .064$) and adding LST-DC above this also did not change the predictive power of the model ($\Delta R^2 = .002$, $p = .624$). Replicating this regression predicting just APM (rather than the *Gf* factor) resulted in a largely identical pattern, except that LST-Basic was a significant, unique contributor in the second model ($\Delta R^2 = .033$, $p = .038$) but remained a non-significant unique predictor in the final model with LST-DC. These two regressions, one predicting *Gf* and one predicting APM, are presented in Table 3.6 and Table 3.7, respectively. Running the regression predicting APM without LST-Basic once again demonstrated that LST-Basic and LST-DC were largely contributing the same variance, as LST-DC became a significant change on its own, above the two WM measures ($\Delta R^2 = .030$, $p = .045$).

Of note, when the same regression was run with SSPAN included as an additional WM measure, the pattern of results changed somewhat. With SSPAN, the addition of LST-Basic became a significant increase in the prediction of *Gf* in the second model; while for the model predicting APM, it was LST-*DC* that was a significant unique predictor in the final model, rather than LST-Basic in the second model. Although, generally, the results of this regression (including SSPAN) are slightly more in line with the hypotheses, the concerns over using the subsample (who completed all tasks) was reason to prioritise the no-SSPAN analyses. This also keeps the approach consistent to that used in Chapter VI, where the same dataset is used to analyse the Swaps task. Nonetheless, because the SSPAN-included version of the regressions was run, in the interest of transparency, these results are available in Appendix A, though we proceed with this chapter with the no-SSPAN results in mind.

Table 3.6. *Full Regression Model Predicting the Gf factor in Experiment 3*

Model	Predictor	B	*t*	*p*	*sr²*	R²	ΔR²
1	Operation Span	.014	0.80	.423	.005	.339	**.339**
	n-back	**.290**	6.22	< .001	.272		
2	Operation Span	.013	0.73	.467	.004	.363	.024
	n-back	**.240**	4.51	< .001	.139		
	LST-Basic	.070	1.88	.064	.024		
3	Operation Span	.014	0.79	.432	.004	.364	.002
	n-back	**.234**	4.28	< .001	.126		
	LST-Basic	.055	1.13	.260	.009		
	LST-DC	.023	0.49	.624	.002		

N=101; bold coefficients p < .05.

Table 3.7. *Full Regression Model Predicting APM in Experiment 3.*

Model	Predictor	B	*t*	*p*	*sr²*	R²	ΔR²
1	Operation Span	.003	0.03	.976	< .001	.285	**.285**
	n-back	**1.331**	5.75	< .001	.251		
2	Operation Span	-.005	-0.06	.950	< .001	.317	**.033**
	n-back	**1.052**	4.00	< .001	.117		
	LST-Basic	**.390**	2.11	.038	.033		
3	Operation Span	.006	0.07	.949	< .001	.323	.006
	n-back	**.999**	3.71	< .001	.101		
	LST-Basic	.253	1.06	.293	.008		
	LST-DC	.207	0.91	.368	.006		

N=101; bold coefficients p < .05.

3.10. Experiment 3: Discussion

The DC effect was again replicated, though once again, there were some differences in this experiment compared to the prior two. When predicting APM alone, the DC effect was clearly replicated, with DC purifying the LST. This was demonstrated by the DC once again subsuming the unique contribution of Basic in the LST regression model, demonstrating that the additional difficulty of the Basic is primarily noise when predicting APM. When using *Gf* rather than just APM (Gf defined through a factor representing both APM and Letter Series), the DC effect was trending but not significant. Unlike Experiment 2, the Letter Series was not significantly correlating with either LST condition, dragging down the overall LST-*Gf* correlation. Although the introduction to Experiment 2 (Section 3.4) argued for this possibility due to the differences in the spatial nature of the Letter Series and APM (where LST is closer to APM in modality), the results of Experiment 2 demonstrated quite clearly that modality was not an issue. In this experiment however, it has become a potential issue, seeing as the Letter Series appeared to have appropriate descriptive statistics (and correlations to tasks other than the LST were also normal). This indicates that the DC effect may be less reliably observed when considering non-visuospatial tasks such as the Letter Series. Given that the DC effect has been observed over three experiments using APM, it does at least appear that the APM is a robust dependent variable for the DC effect to emerge, even if there are some concerns over the overlap in appearance of the tasks.

Another unusual finding was the relationship between the ‘traditional’ WM tasks. The SSPAN and *n*-back displayed strong correlations, both to *Gf* and to each other; yet the OSPAN did not, with weaker correlations across the board. It was expected that the OSPAN would provide a better index of attentional control than SSPAN, given the processing task is more intensive (arithmetic verification as opposed to symmetry judgements). It was argued that the symmetry judgements may be tapping lower-level processing that can be done without requiring a switch of the stored elements to non-active memory. While the SSPAN in this experiment was tainted by always being last in the task set (meaning only the most capable participants completed it), the SSPAN in itself nonetheless had a considerably better showing in Experiments 1 and 3 than the OSPAN in either Experiment 2 or 3, despite this apparent concern with the symmetry judgements. It may be the case that the SSPAN is actually a better measure of attentional control because it primarily measures active storage, rather than some combination of active and passive storage (or switching between the two) that is required in the OSPAN. However, there are reasons to be hesitant over the SSPAN findings in Experiment 3. First, is that the OSPAN has had higher mean performance than the SSPAN in both cases (comparing across Experiments 1 and 2; and within Experiment 3). If the SSPAN did truly require low-level processing, we would expect the mean performance of SSPAN to be higher than the OSPAN, with the more challenging processing. This is a difficult concern to resolve, particularly since the scoring method (total elements rather than total sets), processing threshold (80%+), and the set sizes remained constant across both tasks. The decision to keep the SSPAN for last in the task order was a pragmatic one, over fears that participants would not complete all the tasks in 90 minutes. Rather than risk the participants not reaching one of the more important tasks (like LST) with a fully randomized order of tasks, we instead chose to always put the SSPAN last, so that would be the first sacrificed should participants not reach the end. Although this seemed reasonable at the time, the results now seem to suggest that the subset of 80 participants who actually did complete all tasks may have been qualitatively different from the 26 participants who did not reach the end. That is, the downside of making a "if you get to it" task is that only a certain type of participant "gets to it". Even past the sampling segmentation issue, it is also an issue to have the OSPAN and SSPAN in a fixed order. Given the overlap of the task format between the two, participants approach the SSPAN with the experience of having already completed the OSPAN. Although this does not explain the higher mean scores of the OSPAN compared to the SSPAN, it may have induced a feeling of boredom or fatigue at having to complete such a similar task again, particularly given it would be towards the end of the testing session, where participants may have been in more of a rush to finish. All considered, this issue makes the SSPAN data in Experiment 3 questionable but does not take away from the curious difference observed between the SSPAN and OSPAN across the first two experiments.

Despite these potential issues, it was once again observed that the LST could contribute above-and-beyond the WM measures in predicting *Gf*. This time however, this result was substantially weaker, in part due to the low correlation between LST and Letter Series; and in part due to the remarkably high correlation observed between the *n*-back and *Gf*. In this experiment, the Basic and DC shared most of the variance that was contributed to the *Gf* model. This means that it may have been a contribution of task format (visuospatial). However, using our theoretical explanation of the added load in Basic, it is also possible that the active storage demands of the Basic condition have already been accounted for by the WM measures (mainly *n*-back). Thus, the remaining variance to be predicted is the relational integration that is contributed by both Basic and DC. Overall, the remarkable DC effect has been largely replicated over the course of three experiments, though there were some changes to the strength of the effect and, particularly, to the way it manifests with regards to predicting tasks other than APM and predicting alongside other, varying WM tasks.

3.11. Experiment 4: Introduction

For the final experiment on the LST, we changed the method of analysis to try and determine more specifically what factors contribute to performance. Rather than looking at task-level performance and correlates, the final experiment used an eye tracking approach to determine what parts of the task space were contributing to performance. Time spent gazing at certain areas of each item may indicate the strategies and challenges that participants face when solving LST items. For instance, it is possible that weaker participants simply do not know where to look when solving 4D items. However, if they are looking in the right areas but simply cannot solve the problem, it would indicate that their issue is more based in capacity limits: they simply do not have the binding capacity to engage in the relational integration required to solve the problem. On the other hand, if active storage demands are indeed the issue in solving 2-step items, it is possible that participants will be frequently returning to the interim cell to resolve or remind them of the interim solution.

Thus far, there has been little work on eye tracking in matrix tasks that links areas of interest directly to performance. One study by Laurence, Mecca, Serpa, Martin, and Macedo (2018) found that the only area that seemed to predict performance was the response options (i.e., the options below the matrix that participants choose between as their response). This is not particularly surprising in itself given that participants need to look to their answer before submitting it. However, interestingly, Laurence et al. (2018) found it was not so much the actual time spent on the response options that mattered but rather, the number of ‘toggles’ between the matrix and the response options (i.e., shifting gaze between the matrix and the options). On average, high toggling rates led to a significantly decreased test score.

Although interesting, there are two problems with Laurence et al.’s (2018) findings that Experiment 4 would look to resolve. First, is that the regression analyses were predicting total score rather than item success thus, gaze data and the dependent variable were both aggregates. Although this is less of a problem for their choice of task (Wiener Matrizen-Test 2) because there is less dramatic shifts in difficulty, it is something that needs to be considered for the LST, where items differ in complexity and steps. To solve this issue, the regression analyses will predict item success, rather than aggregate total score. The second issue is that the choice of task (a matrix style task similar to APM) meant that the matrix itself differed substantially between each item – there were little systematic areas of interest. An advantage to using the LST is that each item involves the same rule, and a certain amount of information must always be present (e.g., the target cell, the filled cells involved in the target relation, and any interim cells involved in order to reach the target cell). By identifying these cells in each item, more specific areas of interest (AOIs) can be analysed. The full list of AOIs and the way they are calculated are provided in Section 3.12.1, but an overview is provided here to explain the hypotheses.

3.11.1. Area of Interest hypotheses

Of interest to the hypotheses, the ‘relational cells’ are the filled cells (i.e., contain shapes) with information that must be integrated to come to the outcome of that relation (i.e., the final solution or the interim step solution). ‘Final’ relation cells are the relation cells involved in the final step, while ‘interim’ relation cells are those involved in the interim step (only relevant to 2-step items). It was hypothesized that increased gaze time on both types of relational cells will raise the probability of item success, because these cells are needed to solve the item. If participants cannot identify the cells required to solve the item, then their chance of success falls. ‘Distractor’ cells meanwhile, are filled cells that are not necessary to solve the item. These distractor cells could be empty, and the item can still be solved in the same way. In contrast to relation cells, gaze time on distractor cells indicates that the participant cannot identify the solution pathway and thus, lowers the chance of success. In line with Laurence et al. (2018), it was also hypothesized that higher number of revisits to the response options will also lower item success rate, as it indicates uncertainty. Gaze duration on the ‘final answer response option’ specifically (the option corresponding to the answer) was also predicted to significantly increase item success, because participants need to look here to input their (correct) answer. Although gaze duration on the ‘interim answer response option’ may also relate to item success, it is confounded by participants looking to incorrectly use that option as a response, so no hypotheses are given relating to that metric (but gaze there is likely to inversely relate to success). Item characteristics (RC and Steps) were also included in the analyses, and interactions between gaze data and item characteristics should indicate any differences in predictiveness of the gaze data as a function of item characteristics. It was hypothesized that both these item characteristics (RC and Steps) would contribute strongly to item success, as seen in past studies. Although not interesting in itself in this study, these item characteristics are important to include in the analyses to ensure that the large variance in gaze duration between items is accounted for by item characteristics, leaving only more meaningful variance in gaze duration.

3.12. Experiment 4: Method

3.12.1. Participants

Fifteen participants (nine females) participated in exchange for course credit as part of their first-year summer school program. The average age was 21.94 (SD = 3.30) years. All had normal or corrected-to-normal vision. An additional two participants completed the task, but their data was excluded from the analyses due to concerns over their eye tracking data. One was excluded because they could not pass the calibration test (described in the next section) and one was excluded because their average eye tracking data quality was 74.20%, below the minimum 80% recommended (iMotions Biometric Research Platform, 2018). These participants still completed the LST, but their data is not included here.

3.12.2. Measure and Apparatus

Participants completed the full 36-item version of the LST (Birney et al., 2006). All the items were standard LST items (i.e., Basic items). Participants completed the task using a 14” laptop fitted with an infrared eye tracker (Tobii X2-30) which samples both eyes at 30 Hz. The screen resolution was set to 1366 x 768. Participants sat at whatever range was comfortable to them so long as the eye tracker could pick up their eyes within the reasonable recommend distance of 50cm to 95cm. Participants were calibrated with a standard nine-point procedure to ensure their gaze could be accurately detected within 0.5 degrees. The *iMotions* software (iMotions Biometric Research Platform, 2018) was used to record eye movements and responses to the LST. The average eye tracking data quality was 92.71% (SD = 6.58%). Although this was above the recommended threshold of 80%, the use of an item-based analysis allowed us to specifically exclude trials below this threshold. Overall, 30 of the 540 items were below 80% data quality and were thus excluded from the analyses. After this adjustment, the average data quality was 93.76% (SD = 4.94%).

3.13. Experiment 4: Results

3.13.1. Approach to the analysis

Analyses used participant gaze data within areas of interest (AOIs) as the measure of interest. As seen in the example item in Figure 3.5, the cells of the matrix and the response options were used to create an AOI template applied to all items. The value of each AOI was the total gaze time spent within that AOI on that item for that participant. These values were then transferred to a set of dynamic AOIs, created for each item using an index of item attributes. For instance, for the example in Figure 3.5, item 03, cell 2B is the *final target cell* (the cell with the ‘?’ that must be solved), and so the gaze time value of the *final target cell* is equal to the gaze time value of the cell 2B for that item (whereas for item 04, the ‘?’ cell is in cell 4B, so the final target cell for that item is taken from the gaze time for cell 4B). In addition to *final target cell*, the other dynamic AOIs within the matrix were *distractor cells* (filled cells which have no impact on the solvability of the item), *final relation cells* (filled cells involved in the relation of the final step), *interim target cell* (the cell that must be solved in the first step of a 2-step item, i.e., the interim step), and *interim relation cells* (filled cells involved in the relation of the interim step). In addition, two dynamic AOIs corresponding to the response options were calculated: *final-answer-RO* (the response option with the answer to the item) and *interim-answer-RO* (the response option with the answer to the interim step). For *distractor cells, interim relation cells*, and *final relation cells* (all of which may have more than one cell per item), the value was a sum of all the cells that corresponded to that attribute for that item.

The target, relational, and interim cells were derived from Birney et al.’s (2006) RC analysis of the LST, though an additional item analysis was then conducted for each item to determine if there were alternate solution pathways. For some items, there were indeed multiple solution pathways, which made calculating distractor and relation cells difficult. For these, we first assumed that the most relationally simple pathway was taken. In the event of a tie (e.g., an item where two binary solution pathways were available), each separate solution pathway was calculated separately, and the final value of the relation cells was equal to the *highest* gaze duration solution pathway used by each participant. For distractor cells, the cells were only summed if they did not contribute to *any* potential solution pathways. In other words, distractor cells were filled cells that, if removed and turned to empty cells, would not affect the solvability of the item *regardless* of the pathway taken. Although this approach may result in some loss of gaze data if participants switch solution pathways through the problem, it was the most straightforward solution to ensuring there was only one set of relation and distractor cells per item for use in the analyses.

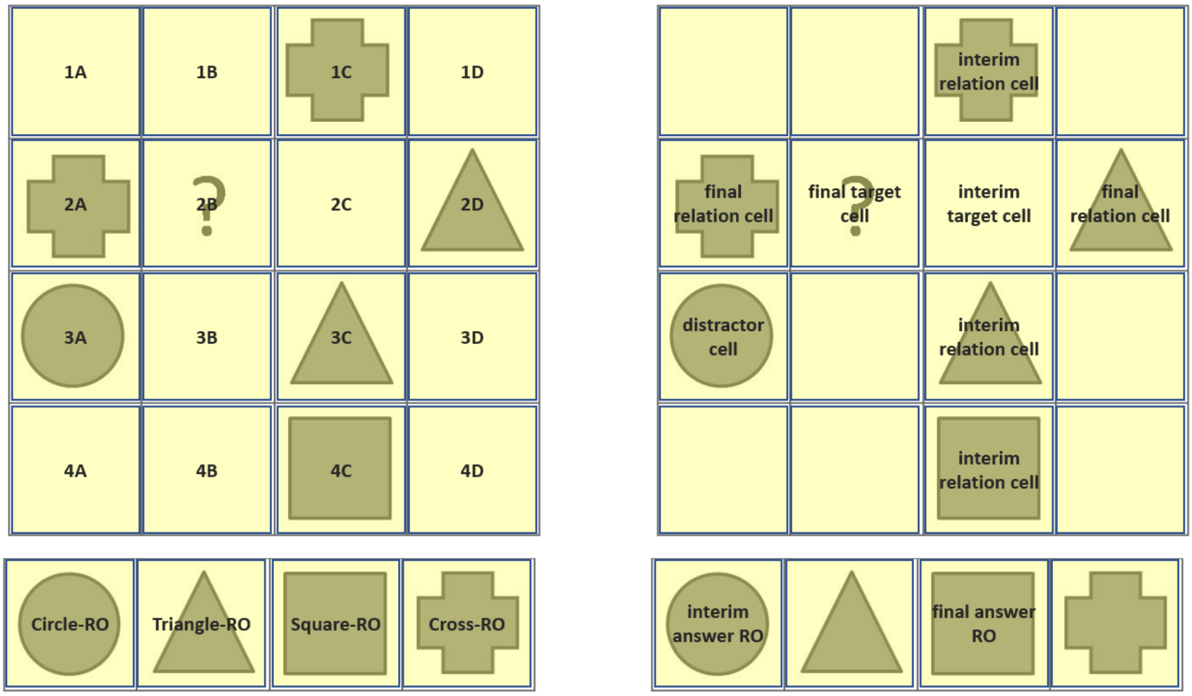


Figure 3.5. Areas of interest (AOIs) for the LST. The matrix on the left displays the AOI template analogously applied to all items. These template AOIs are converted to dynamic AOIs for each item, as demonstrated by the example matrix on the right. For 1-step items, there is no *interim target cell, interim relation cells*, or *interim answer RO* (response option). *Distractor cells* are shape-filled cells that have no impact on the solvability of the item (i.e., they could be turned to empty cells and the item would have the same solution pathway). For items with multiple paths to solution (e.g., two sets of relation cells per target cell), the set of relation cells with the highest amount of gaze duration (per participant) are recorded as the relation cells for that item.

Finally, we also included *RO-revisits*, a measure of the number of times a participant returned to the response options on each item. Although not a measure of gaze duration, revisits is nonetheless a gaze metric, one which Laurence et al. (2018) found was the best predictor of test scores on a similar, matrix-style reasoning task.

The program recorded gaze duration data in milliseconds, but values are reported in seconds for interpretability. Hypotheses were tested using binary logistic regression on item-level data, using item metrics (*RC*, *steps*) and gaze metrics (e.g., *final target cell, final relation cells, RO-revisits*, etc.) for each item predicting success on that item (0 for incorrect, 1 for correct). Coefficients for each predictor were recorded and evaluated statistically by the change in log-odds. Confidence intervals for odds ratios are reported, for ease of interpretability (CIs containing 1 indicate non-significance).

3.13.2. Gaze time descriptives and logistic regressions

Overall, performance was similar to that described in the earlier experiments for RC (2D M = .94, SD = .24; 3D M = .83, SD = .37; 4D M = .58, SD = .50) and Steps (1S M = .82, SD = .38; 2S M = .75, SD = .43). Descriptives for gaze metrics are provided in Table 3.8. These mean values demonstrate that, on average, about 3.5 seconds were spent on final relation cells of each item, while 4.9 seconds were spent on interim relation cells. The high variance in these descriptives is to be expected, considering they average across item types.

Table 3.8. *Gaze time metric descriptives.*

	Mean	SD
Final answer RO	0.97s	0.74s
Interim answer RO (2S only)	0.67s	0.85s
Final target cell	2.21s	3.22s
Interim target cell (2S only)	1.75s	2.19s
Final relation cells	3.54s	4.00s
Interim relation cells (2S only)	4.90s	5.29s
Distractor cells	1.87s	2.96s
RO Revisits	4.22	5.32

N = 510 (15 x 36) item responses (255 for 2S only metrics)

For the first regression, item success was predicted using *RC, Steps, final answer RO, final target cell, final relation cells, distractor cells,* and *RO revisits*. As hypothesized, *RC* was a significant predictor of item success (CI95% = [0.172, 0.400], p < .001), as was *Steps* (CI95% = [0.216, 0.802], p = .009), both lowering the chance of success with increases. For the gaze metrics, *final answer RO* was a significant and very powerful positive predictor of success (CI95% = [17.77, 90.95], p < .001), though this was unsurprising, as it was attributable to the fact that participants needed to input their answer by clicking the corresponding response option. *Final target cell* was also significant (CI95% = [0.801, 0.983], p = .022), though in a negative direction: for every 1 second spent looking at the *final target cell,* there was, on average, a 12.3% reduction in the chance of correctly answering the item. The *distractor cells* were also significant (CI95% = [0.750, 0.935], p = .002) in a negative direction: for every 1 second spent looking at distractor cells, there was, on average, a 16.3% reduction in the chance of correctly answering the item. The number of *RO revisits* (toggling rate) was also significant (CI95% = [0.736, 0.867], p < .001) in a negative direction: for every additional revisit to the response options, there was, on average, a 20.1% reduction in the chance to solve the item correctly. Contrary to the hypothesis, the *final relation cells* were not significant predictors of item success (CI95% = [0.966, 1.112], p = .001). Table 3.9 displays the full output of this regression.

Table 3.9. *Output of Binary Logistic Regression with Item Characteristics, Gaze Time on Areas of Interests (AOIs), and Revisit Rates predicting Item Success (1S and 2S items).*

	Exp(B)	CI-Exp(B)	Sig.
Relational Complexity	0.263	0.172, 0.400	**< 0.001**
Steps	0.416	0.216, 0.802	**0.009**
Final-answer Response Option (sec)	40.207	17.774, 90.952	**< 0.001**
Final-answer Target Cell (sec)	0.887	0.801, 0.983	**0.022**
Final relation cells (sec)	1.036	0.966, 1.112	0.323
Distractor cells (sec)	0.837	0.750, 0.935	**0.002**
Response Option Revisits (#)	0.799	0.736, 0.867	**< 0.001**
Constant	*329.65*		***< 0.001***
χ^2 =240.17, df = 7, p < .001			
Classification Accuracy = 88.8%			
Nagelkerke R^2 = .582			

N = 510 items

The second regression included the same predictors as above, but also added interim gaze metrics as additional predictors (*interim answer RO, interim target cell, interim relation cells*). Because interim gaze metrics were only calculated for 2S items, only 2S items were included. This regression was conducted over two models. The first model aimed to replicate the results of the first regression (i.e., interim AOIs were not included), while the second model added the interim AOI metrics. The first model mostly replicated the previous regression. However, this time, the *final-answer target cell* was not a significant predictor of item success, (CI95% = [0.748, 1.090], p = .288); but the *final relation cells* were (CI95% = [1.005, 1.416], p = .044), such that for every 1 additional second spent looking at the final relation cells, there was, on average, a 19.3% increase in the chance of solving the item correctly. In the second model, the pattern of predictions for the previous predictors remained the same. Of the three new predictors, only *interim answer RO* was a significant predictor, in a negative direction (CI95% = [0.108, .813], p = .018. However, as with the other response option AOIs, this should be interpreted with caution, since those looking to input their answer look towards the response options (in this case, inputting the interim response would result in an incorrect answer, so the chance of success decreases). Contrary to hypotheses, the other two predictors, *interim target cell* and *interim relation cells* were not significant predictors, *p*'s > .05.

3.14. Experiment 4: Discussion

The overall purpose of Experiment 4 was to further elucidate the processes involved in successful performance on the Latin Square Task. As it turns out, most of the AOI predictors ended up being related to *un*successful performance, with longer gaze on certain metrics related to higher failure rates. The predicted impact of RC and Steps was found, with a large detriment as these item characteristics increased. This is not at all surprising given the earlier experiments on the LST but was nonetheless necessary to account for what would otherwise be noise in our more subtle analysis of the eye tracking metrics. Of the gaze metrics, we found the same toggling findings as Laurence et al. (2018), where the number of revisits to the response options was inversely related to success. This may indicate that unsuccessful participants are more unsure of their answer, frequently returning to the response options to search for possible solutions. Successful participants meanwhile, understand the rules, and are only looking to the response options to confirm their response.

The analyses on gaze duration metrics were more novel, with no work (until now) being done that specifically identifies areas of interest related to success on the LST. Of particular interest was the relational cells, which are necessary to solve the item. These were, contrary to the hypotheses, not related to item success overall. This could be because the clear, singular rule of the LST (that each row and column must contain only one of each element) and the presence of the '?' in the target cell makes it easy for all participants to eventually identify the relational cells, though only successful participants then know what to do with these important cells. Although the relational cells become less obvious at the higher complexity of 4D items, there are often fewer filled cells in general in 4D items, so identifying the appropriate cells is still common to all participants. This conclusion is somewhat incomplete, particularly when considering the important findings on the distractor cells. Distractor cells are cells that are filled with shapes but are in no way necessary to solve the item. Perhaps the most insightful finding in this experiment is that gaze time on these distractor cells was significantly negatively related to item success, with (on average) every 1 second of gaze time spent on distractor cells leading to a substantial 16.3% decrease in the chance of solving the item correctly (when controlling for item characteristics and other AOIs). Taken in isolation, the distractor cells finding seems to indicate that success may indeed be related to identifying the important cells (relational cells) among the filled cells. However, taken together with the non-significant relational cells finding, the results suggest that all participants eventually identify the important cells, but those who linger on the distractor cells are the ones that fail. Successful participants can identify the important cells and swiftly disregard the distractors, making their solution process more effective. Failing participants can also identify the important cells, but they appear to have more trouble disregarding the distracting information given in the matrix.

The results of the interim steps on 2-step items were also insightful. The fact that neither the interim target cell nor the interim relation cells predicted success seems to indicate that failing participants could identify the initial step despite the target ‘?’ not helping them with this identification. In this case though, the final relation cells did predict success. This is particularly interesting since the interim target cell is not marked, so if participants were struggling to find the solution pathway, the gaze data on this cell would indicate this is where unsuccessful participants are getting stuck. However, it appears that successful participants are the ones that not just identify the interim cell, but successfully move on from it to the second step. Either that, or they are working backwards – using the ‘?’ to identify what cells are required to solve the item. Whichever it is, it is something unsuccessful participants are failing to do. Of course, once again, the distractor cells also related negatively to success.

Experiment 4 provided another perspective from which to analyse the LST. Analysing eye tracking results necessarily involves some assumptions about the data being made. For instance, although it appears distractor cells are causing problems for failing participants, it is possible that failing participants are simply looking all over the matrix. This could also explain why relational cells were not significantly related to success – both successful and unsuccessful participants look at the relational cells, but for different reasons: successful participants identify the relational cells are integral to the solution while unsuccessful participants are simply looking everywhere. Thus, it may not necessarily be the distractor cells causing them issues but rather, gaze time on distractor cells are an outcome of their poor capability to solve the item. It should also be said that the data here is limited to a small sample size, and how these gaze metrics relate to individual differences variables, such as performance on Gf tasks, would be of interest in future research. Nonetheless, these eye tracking results are (to the best of my knowledge) the first to be conducted on the LST and certainly the breadth of metrics that can be considered in the LST indicate it may be a fruitful task for future gaze analysis (in comparison to Laurence et al. (2018) who was largely limited to just response option revisits due to the choice of task). In addition, they certainly contribute additional insight into the DC effect found throughout Experiments 1-3, which we turn to in the next section.

3.15. General Discussion

The overall goal of this first chapter was to replicate Bateman's (2015) remarkable DC finding and discover more about how and why it manifests. The DC effect is a phenomenon observed in the LST where a 'dynamic completion' version of the task which minimizes passive storage demands both (unsurprisingly) increases average performance on the task and (surprisingly) *increases* the correlation with more complex abstract reasoning (Gf) tasks such as Raven's matrices (APM). Over the course of three experiments, the DC effect was largely replicated each time, with some differences in the strength of the effect and how it manifested. Table 3.10 provides a summary of the major findings of these first three experiments. As seen in rows 8-9 of Table 3.10, the DC effect was clearly replicated across all three studies in that LST-DC (row 9) had a strong unique contribution over-and-above LST-Basic (row 8). There were some differences with how this manifested across the three studies. In Experiment 2, LST-Basic retained a unique contribution. Experiment 2 was also the only study where LST-Basic was a stronger unique predictor of Gf than LST-DC. Differences between the findings also emerged when controlling for WM (rows 10-11), though this is not entirely surprising given the change in WM criterion tasks throughout the experiments: generally, the DC effect (controlling for WM) got weaker as the WM task set became more comprehensive.

Table 3.10. *Summary of the major findings with the LST in Experiments 1-3.*

			p		
			Study 1 (*n* = 125)	Study 2 (*n* = 100)	Study 3 (*n* = 106)
1	ANOVA	Linear trend for RC (accuracy)	< .001	< .001	< .001
2	ANOVA	Linear trend for Steps (accuracy)	< .001	< .001	< .001
3	ANOVA	DC accuracy higher than Basic	< .001	.003	< .001
4	ANOVA	Interaction of RC x Condition	.381	-	.047
5	ANOVA	Interaction of Steps x Condition	.019	-	.568
6	ANCOVA	Linear RC covaries with *Gf*	.007	< .001*	.006
7	ANCOVA	Linear Steps covaries with *Gf*	.014	< .001*	.022
8	Regression	LST-Basic unique contribution in final model predicting *Gf* (LST only)	.137	< .001	.092
9	Regression	LST-DC unique contribution in final model predicting *Gf* (LST only)	.012	.020	.012
10	Regression	LST-Basic unique contribution in final model predicting *Gf* (controlling for WM)	.362	.003	.260
11	Regression	LST-DC unique contribution in final model predicting *Gf* (controlling for WM)	.008	.061	.624

ANOVA *p*-values (rows 1-5) are significance tests of *F* values from ANOVA comparing accuracy; ANCOVA *p*-values (rows 6-7) are significance tests of *F* values from ANCOVA comparing accuracy with APM/Gf added as a covariate; Regression *p*-values (rows 8-11) are significance tests of the semi-partial correlations (sr^2) of the predictor (LST-Basic / LST-DC) in the final model, either with LST only (rows 8-9) or with WM criterion measures controlled for (SSPAN/OSPAN/*n*-back; rows 10-11). Significant *p*-values are highlighted in green for ease of comparison between studies.

RC = Relational Complexity; DC = Dynamic Completion; APM = Raven's Advanced Progressive Matrices (20 item); *Gf* = APM (Experiment 1) or latent variable extracted from APM and Letter Series (Experiments 2+3).

* These *p*-values are derived from separate two separate ANCOVAs with either RC or Steps (i.e., not both together). The increased item size (and increased range in potential scores) leads to an overestimation of the covariance effect in Experiment 2, relative to Experiments 1 and 3.

In Experiment 1, the DC effect was clearly replicated in predicting APM, and this effect persisted despite the addition of classic WM tasks such as the complex-span being added to the predictive model. The shared variance between the Basic LST and complex-span (not seen in the DC LST) demonstrated that these tasks do indeed share components which is theorized to be an active storage demand related to attentional control. The fact that the DC condition predicts APM over-and-above this active storage component demonstrates that attentional control is not integral to tapping into APM; and the fact that the correlation actually *increases* indicates that these attentional control components may only serve to add noise (i.e., residual variance) to predicting APM. The DC effect “purifies” the task, isolating only the most important components – theorized to be related to relational integration. In Experiment 2, there was once again a strong correlation between DC and Gf, though this time, it was not higher than Basic. Nonetheless, the addition of the Letter Series to the Gf factor and the *n*-back to the WM factor helped to confirm that the DC effect largely persisted beyond consideration of merely APM. It was somewhat unfortunate then, that the DC effect was unexpectedly weakened by the addition of the Letter Series in Experiment 3 (which did not occur for Experiment 2). This indicates that the DC effect may not be completely robust to changes in the criterion tasks, though it nonetheless remained a remarkable finding given what DC does to the difficulty of the LST.

Unfortunately, the DC result was not completely straightforward to interpret, due to the inconsistent interaction with Steps. The DC effect and Steps effect are in contrast to one another, with one loading on active memory (Steps) and the other reducing the impact of active memory (DC). It was thought that DC was acting on the LST by reducing the impact of the Steps manipulation, but this does not seem to be the case, as the DC effect did not always show an interaction with Steps. That is, DC was often impacting on 1-step items as well as 2-step, even though only 2-step should show the benefit if the DC is as simple as a reversal of the additional load incurred by 2-step.

Unexpectedly, the linear RC effect (RC covarying with Gf) did emerge, though once again, this seemed to largely be a result of a qualitative difference between 2D and 4D items rather than a perfect linear effect going from 2D to 3D to 4D. This was evident because, when placed in a regression, 2D and 4D items tended to predict unique variance in Gf while 3D items consistently did not. That is, the 2D and 3D items seem to predict the same share of variance in Gf, separate from 4D. Thus, although it was theorized that the LST and Gf primarily share variance through relational integration, it cannot simply be measured by varying relational complexity demands. This is because (at least in adult populations), RC does not manifest as a linear increase in demands but rather, a qualitative difference between 4D items and items of lower complexities. Although the overall linear trend is significant and would indicate this to be the case, the linear effect is driven completely by 4D items undergoing a substantial drop in performance rather than a smooth linear performance decline across the three levels. This could be because 2D and 3D items can be solved by simply applying a sequential *shape-based* application of the fundamental rule (each row and column may have only one of each shape), adding each different type of shape (e.g., square, circle, triangle...) to the running list of integrated shapes and leaving only one shape left that must be the answer (cross). 4D items meanwhile, cannot be solved using this sequential approach and instead, must be solved by considering only one type of shape (e.g., cross) and applying a *dimension-based* approach focusing on the indirect interpretation of the fundamental rule (each row and column *must* have one of each shape). Although the differences in approach means there may be a strategic component to success at the LST (which is to be explored), the two approaches (shape-based vs. dimension-based) are nonetheless also different in relational integration demands, at least in theory (Birney, 2002). This is because, in the shape-based approach, the elements in question (shapes) can be systematically chunked while the dimensions (rows and columns) cannot (Birney, 2002). Thus, the DC effect may manifest because it allows items that would normally be restricted to dimension-based approaches to also be solved using shape-based approaches. For example, consider the items in Figure 3.6 (from Bateman, 2015). Item 43 can be solved *either* with a single 4D step using a dimension-based approach (left of Figure 3.6) *or* using a long, multi-step pathway that heavily loads on storage demands but does *not* require levels of relational complexity above 3D (middle of Figure 3.6). In contrast, item 46 (right of Figure 3.6) cannot be solved in any way except for a single 4D step using a dimension-based approach. Although there is overall less variance when employing DC (average performance is closer to ceiling and standard deviation decreases), the variance that remains is *only* variance resulting from the high relational integration demands involved in the dimension-based approach, as in item 46. Item 43 meanwhile, is introducing unhelpful noise because participants can be solving it either using a dimension-based approach (loading heavily on relational integration and related to Gf) *or* using a shape-based approach (loading heavily on active storage and unrelated to Gf). DC is effectively eliminating this residual variance and maximizing the contribution of only the most important items to Gf, like Item 46.

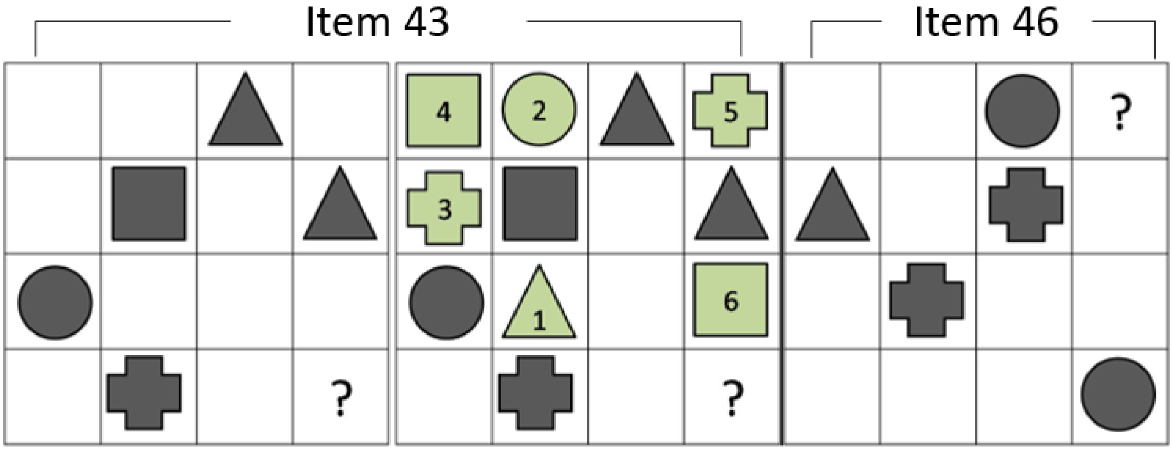


Figure 3.6. Taken from Bateman (2015). Example of how dynamic completion enables intricate, multi-step ‘shape-based’ solutions not normally intended for the LST. Item 43 (left and middle) is a 4D1S item intended to be solved with one ‘dimension-based’ 4D step, but DC enables a long 6-step chain of simpler relations. Item 46 (right) meanwhile, is also a 4D1S item but can only be solved with one 4D step and DC does not enable any additional solution pathways.

A limitation on this explanation is that we cannot conclusively say whether the failure to employ dimension-based approaches is due to limitations in binding capacity or simply because failing participants cannot identify that the dimension-based strategy is required. An experiment could help solve this by inducing strategies through the instruction. For instance, highlighting a change in the fundamental rule from “each row and column *may* contain only one of each shape” to clearly specifying “each row and each column *must* contain one of each shape” could be enough to induce participants towards a dimension-based approach. The exact wording of the rules is something that would need to be determined, as would the presence, contiguity, and quantity of reminders of these rules. Although this suggested experiment would be ideal for answering the strategy question, there are novel results from eye tracking that can support this conclusion.

In Experiment 4, a different approach – eye tracking – was used to attempt to discover how the Basic task is solved. It was expected that (because of DC), time spent on the cells involved in the target relation could be related to DC. It turned out the results were not this straightforward. Gaze time on target relation cells had no relationship with item performance, which could be because (a) even unsuccessful participants can identify the target relation cells, or (b) successful participants tend to be more efficient and do not require long gaze time on the target relation cells. More insightfully, time spent on *distractor* cells (filled cells not relevant to the solution) was inversely related to item success. Although all participants can identify the required relational cells, successful participants are better at maintaining focus on these and resisting distraction from irrelevant cells. This is further supported by toggling rates (revisits to response options) also being inversely related to success. Together, these results indicate that more successful participants are better at sticking to goal-relevant information and are more efficient at coming to the solution – they know their answer before they look at the response options, and only look to the response options to confirm their already-known answer. Unsuccessful participants, meanwhile, are distracted by goal-irrelevant information and tend to use the response options to try to work out the answer, either as a process of elimination or in order to remind them of the shapes involved in the relation. This latter possibility means unsuccessful participants may be less aware of the task rules or shapes involved (requiring reminders of the shapes in each set) but seems less likely considering they also identify the correct relational cells within the matrix itself.

Although goal orientation is related to attentional control (Kane et al., 2004), the goal orientation used here is more strategic in nature rather than a vigilance aspect commonly considered in attentional control theories. That is, the results of the eye tracking supplement the DC findings by demonstrating that unsuccessful participants are looking for other solution pathways. This is either because they are unable to identify the dimension-based approach or because they lack the binding capacity to engage in the high-complexity dimension-based approach; or even, potentially, as some combination of both in recognition of their low binding capacity which would implicate a metacognitive strategy component that can also be related to their choice of solution pathway. In any case, this search for other solution pathways leads to unsuccessful participants stumbling upon the distractor cells more often (as we observed in the gaze analysis). Sometimes these alternate solution pathways work and sometimes they do not, contributing noise to the Basic LST. The DC effect manifests primarily by magnifying the effect of items with only a single high-complexity dimension-based solution. Although it would have been ideal to compliment the eye tracking results with criterion tasks (Experiment 4 lacked the sample to do this), the results nonetheless give some indication of why DC works. Taken together, these results also have practical applications for the future use of the LST, demonstrating that factors beyond just RC and Steps are important for deploying the LST. Factors such as the ratio of filled cells to empty cells[8] and the ease of alternate, shape-based approaches to high-complexity items are factors that must also be considered when employing the LST, particularly when trying to relate the task to Gf. If these explanations of the DC and eye tracking data is correct, then they would predict that a version of the task which completely removes distractor cells may be just as

[8] The number of empty cells was considered briefly by Birney (2002, p.95). It did not seem to impact on item difficulty though there was a slight increase in response times as a result of fewer empty cells. Simply given the increase in raw visual information required to process (more shapes fill the matrix), a slight delay in response times is unsurprising. However, given the subtlety with which DC manifests (primarily through a select few items, such as Item 46 illustrated in Figure 3.6), to truly conclude on the impact of distractor cells may require more specific item-level breakdowns rather than aggregates.

effective as DC at isolating the relational integration effect in high-complexity items through reduction of storage-related noise. This version would also not require the additional instructions and complex task coding and scoring that was required by the DC condition.

3.15.1. Limitations

Some limitations are worth discussing. Obviously some of the explanations made above are contingent on assumptions made somewhat indirectly from the current data. For instance, the two approaches (shape- vs. dimension-based) have not been directly contrasted in an experiment, though there is some evidence from verbal protocol data that high-performing participants tend to look for the most efficient solution pathway rather than simply trying to solve it any way they can (Birney, n.d.-b). Another issue was the relative lack of items in each condition: Basic and DC each only had the same 12 items representing them in each experiment. Although these items were chosen randomly (Bateman, 2015) from the original set of 36 (Birney et al., 2006), it is plausible that the DC items chosen just happen to be ones that best capture variance in Gf, even in spite of the DC manipulation. Although unlikely, the low number of items (12) means this is a possibility, since the effects are less likely to average out. Replicating the study using additional items in each set, or by using different items in each set, or even simply inversing the items used in Basic and DC could resolve this concern by determining that the DC effect is not tied to the 12 items being deployed. Although not a methodological limitation *per se*, a caution that is worth repeating is that the DC effect was not consistently replicated when predicting Letter Series, a non-visuospatial Gf task. In Experiment 2, DC effortlessly predicted Letter Series, with little difference between whether APM or Letter Series was used as a Gf measure; while in Experiment 3, it was clear that the DC effect only emerged when predicting APM and not Letter Series. Although even generally the DC effect should look to be replicated with additional Gf tasks, the very fact that a non-visuospatial dependent task can produce such inconsistent results highlights a need to reconsider the DC effect beyond just predicting APM. Beyond criterion tasks, it would also be of interest to consider DC in tasks other than the LST. However, part of the theorizing on why it works so well for the LST is because the LST is a task based on relational integration – something that the DC condition does not take away. It is difficult to find tasks that are founded so fundamentally in relational integration where the storage demands can be so elegantly removed. One example of this was seen in the original demonstration of 'monitoring' tasks by Oberauer et al. (2008), who compared storage vs. no-storage versions of the monitoring tasks. Much like the current study, Oberauer et al. found that non-storage versions of monitoring tasks were equally as good at predicting Gf as storage-loaded versions.

3.15.2. Conclusion

Despite the cautions mentioned above, the DC effect nonetheless remains an important phenomenon in the LST, going against conventional wisdom that the more difficult a task is, the more it should relate to advanced higher-order constructs like Gf (Stankov, 1993). A task that becomes more difficult may increase the correlation with Gf, but this is manifesting through psychometric properties, such as an increased range removing ceiling effects and better capturing the full range of abilities in the population. What the current study has repeatedly demonstrated is that if a task becomes more difficult through the increase in load on components unrelated to Gf (such as storage), it will *decrease* the correlation, and this is *in spite* of the otherwise better psychometric properties like reducing the ceiling effect. This is because the increased range is only a result of increased noise. The current experiment demonstrates that careful consideration of the theoretical components underlying a task should be prioritised before optimising the psychometric properties of the task.

IV. STUDY 2: THE ARITHMETIC CHAIN TASK

This chapter is published in Bateman and Birney (2019). [Bateman, J. E., & Birney, D. P. (2019). The link between working memory and fluid intelligence is dependent on flexible bindings, not systematic access or passive retention. *Acta Psychologica, 199, 102893*]. There are minor changes to terminology and flow to fit the thesis.

In Chapter II, the growth of working memory (WM) theories was discussed, evolving from traditional views of a multi-componential system, comprised of static 'slave' stores with an overarching executive processor (Baddeley & Hitch, 1974) to an attentional system which provides temporary access to representations within memory based on their level of activation (Baddeley, 1993; Oberauer, 2009a; Shipstead et al., 2016). Relational theories such as Oberauer et al.'s (2007) see this temporary activation as based on bindings, where elements are bound to schema within a coordinative system that conveys relations among the elements. In this way, the 'capacity' of WM is not dictated by the raw number of elements that can be stored but rather, by the number of bindings that can be simultaneously established. The binding view has implications for the often-reported link between WM and Gf (Ackerman et al., 2005), where the intricacies underlying this relationship are not fully understood. The current study utilizes the Arithmetic Chain Task (ACT) from Oberauer, Demmrich, Mayr, and Kliegl (2001) in order to determine task factors associated with WM performance and the link to Gf. We replicate the Oberauer et al. (2001) findings that actively accessing previously stored information during processing impacts processing performance, while passively storing unrelated information does not. However, we also extend these findings by considering the effect of systematic chunking of this stored material. We find that a condition which prevents systematic chunking (by forcing the access of stored bindings in a random order) is critical to linking the ACT with a composite measure of the shared WM and Gf variance defined by prototypical tasks.

Consistent with Oberauer et al.’s (2007) concentric model (also called the ‘three-embedded-component’ model), the recent work of Chuderski (2014), and the theorising of Shipstead et al. (2016), rather than seeing WM as a subordinate process of Gf, we begin with the premise that there is a mechanism (or set of mechanisms) that is common to both *WM* and *Gf*. Like Oberauer (2009a), we see this mechanism as having the ability to dynamically establish and maintain bindings within WM. Our view is that capacity is dictated by the strength and flexibility of these bindings. In this way, WM is not simply about storing static information and executing processes on the information system. Rather, WM is a coordinative system that provides access to information by binding elements or chunks of elements (Cowan, 2010) to positions within a relational schematic (Oberauer, 2009a). For instance, recalling previously listed letters such as A-J-L may involve constructing a relation of the running sequence where A is bound to position 1, J to position 2, and L to position 3. These bindings are held within a region of ‘direct access’, where central attention (the focus of attention) can be redirected between bindings (Oberauer, 2009a). Because elements must be bound to be represented in the direct access region, active elements are necessarily connected by some common relation (Cowan, 2001); in this example, the temporal order of the running sequence. Recall and recognition tend to be superior when the common connecting relation is intuitive (i.e., the elements are naturally grouped by form or position in a display or by semantic similarity in a list) but this also means WM can be easily fooled if changes are made to the display or sequence that maintain the active relation (Roediger & McDermott, 1995). Representations outside the direct access region are relegated to long-term memory, but can be more or less easily brought into the direct access region based on the level of associative activation, akin to connectionist models (Anderson, 1983; Collins & Loftus, 1975).

A binding approach such as Oberauer's (2009) changes how we consider traditional *WM* tasks. For instance, the complex-span approach (e.g., Daneman & Carpenter, 1980) is seen as a hallmark WM assessment paradigm (Unsworth & Engle, 2007b). Participants alternate between storing some element (e.g., a letter) and a simple processing task (e.g., verifying an arithmetic problem) before being asked to recall the sequence of elements. In modular views of WM where storage and processing are clearly segmented, it makes theoretical sense that the task procedure is also segmented. The storage component of the task represents the capacity while the processing is there to occupy and distract central attention in order to prevent active rehearsal. When considering a capacity based on bindings, the alternating nature of the complex-span task becomes theoretically more insightful because the distracting processing disrupts the building and maintenance of the relation representing the running sequence. Distractors must be encoded and active to be processed, but they must also be kept distinct from the relation of the primary running sequence, which is easier if the distractors are of a different modality (Oberauer, Farrell, Jarrold, Pasiecznik, & Greaves, 2012). Good performance is thus about establishing strong bindings that persist despite central attention being frequently drawn to auxiliary processing; and about efficiently dissolving the unrelated bindings involved in the auxiliary processing once the processing is complete.

In the current experiment, we aim to contribute to understanding of capacity limits in the binding system by comparing (a) the relevancy of the auxiliary task to the primary task; (b) the *systematicity* of the link between the auxiliary task to the primary task; and (c) what effects these task manipulations have on the relationship to a prototypical WM task: the complex-span; and a prototypical Gf task: Raven’s Progressive Matrices (J. Raven, 1989); as well as a common factor that represents their shared components. In the remainder of the introduction, we outline our choice for the Arithmetic Chain Task and explain the task manipulations.

4.1. Introduction to the Arithmetic Chain Task

Interesting results have emerged from research investigating the link between the auxiliary task and the primary task in *WM* paradigms. Oberauer et al. (2001) explored the difference between *access* to stored information and mere passive *retention* of stored information on arithmetic processing. Using the Arithmetic Chain Task (ACT), Oberauer et al. found that processing was impaired by a storage load only when the stored information had to be accessed and used as part of the processing. In the ACT, participants are shown, operator by operator, a simple arithmetic equation involving a number of digits, three of which have been replaced by the letters, XYZ (e.g., 5 + 7 – X + Y + 3 + Z – 4). In the *control* condition, a key showing the numeric values of X, Y, and Z is displayed above the equation (e.g., X=2; Y=4; Z=1). In the *retention* condition, prior to the arithmetic phase, participants are also briefly shown to-be-remembered variable-value mappings associated with different letters, ABC (e.g., A=4, B=3, C=7). These ABC mappings were not relevant to the arithmetic but must be recalled after the equation is solved. The retention condition is otherwise identical to the control condition with the XYZ mappings displayed above the equation. In the *access* condition, the direct numeric XYZ mappings were not displayed on the screen, instead they were displayed indirectly as additional mappings to the previously presented ABC mappings (e.g., X=B, Y=C, Z=A). In this way, the only difference between the retention and access conditions is whether the stored information must be accessed during the arithmetic processing. It is the *access* condition that is the focus of our investigations.

Oberauer et al. (2001) predicted performance declines in the access condition only, because the ABC load cannot be relegated to long-term memory for the duration of the arithmetic. Instead, the load must be kept active in the region of direct access. Indeed, Oberauer et al. found that arithmetic performance was only significantly degraded when the XYZ mappings had to be accessed through the task-critical ABC load as in the *access* condition, indicating that maintaining direct accessibility to the bindings was a detriment to active processing, as compared to passively retaining the task-irrelevant ABC load in the *retention* condition. However, because the XYZ variables were mapped to ABC in a random order in the *access* condition, it is possible there was a significantly increased demand in having to restructure the mappings for use in the processing. That is, in the *retention* condition, the digit-to-ABC mappings can be systematically chunked and retained with only one binding (ABC = 437), while the XYZ mappings are available at all times during the arithmetic (X = 2 / Y = 4 / Z = 1). Conversely, in the *access* condition, the ABC mappings cannot be chunked with only binding (as in ABC = 437) because the relational information defining each separate mapping is necessary in order to independently match them to the randomly-ordered XYZ mappings (e.g., X = B / Y = C / Z = A). If the ABC mappings are chunked into one binding as in the retention condition (ABC = 437), the relational information defining each separate variable-value mapping (e.g., A = 4) is rendered inaccessible (Halford et al., 1998), as only the entire chunk is accessible when represented as such in WM (ABC = 437). In practice, this means the strategy of storing a single binding (ABC=437) breaks down when randomly-ordered, because the ABC mappings cannot be systematically applied to XYZ (XYZ=ABC=437) as the ordering of the chunk changes (randomly) with every new item (e.g., X = B, Y = C, Z = A).

An ACT access item with randomly-ordered mappings requires independent relational information for each mapping or at least, the rearrangement of the order of the chunk during the problem (as the ABC-XYZ mapping orders are only apparent once the problem begins). This requires either three separate bindings to be active or the ability to retrieve and rearrange the bindings, respectively. Thus, the additional cost incurred may not be due to access over retention, but due to the additional demands placed on binding due to the inability to systematically chunk the variable mappings into an easily accessible, ordered form. If the increased demand of access only emerges in a *random-access* version (e.g., ABC=YZX), as opposed to a *fixed-access* version (i.e., ABC=XYZ), it would indicate performance on the task relies on the flexibility of the bindings during processing, rather than merely accessing the stored information.

Thus, in the current study, the format for control, load, and access conditions were adopted but in addition, we consider two types of access: *fixed* (i.e., ABC=XYZ) and *random* (e.g., ABC=YZX). The following task analysis fully explains our rationale and how systematic ordering *(systematicity)* can bypass binding demands of the fixed-access variant of the task, but not the random-access variant.

In the ACT, the primary demands are in the construction of arithmetic relations such as addition (e.g., 2 + 3 = 5). *Addition* is a relation formed through the binding of three variables within a schema: two addends (2 and 3) and a sum (5), and is explicitly classified as a ternary relation in Halford, Wilson, and Philips' (1998) Relational Complexity (RC) scheme (p. 808). Where the sum would normally require derivation, the simple arithmetic of single-digits can be bypassed using the knowledge of over-learnt relationships acquired early in schooling (e.g., between |2,3| and |5|). In this way, the *effective complexity* of single-digit addition can be systematically chunked into a binary relation.

Processing in the control condition of the ACT thus amounts to analogous instantiations of a series of such binary relations, one for each operand. Additional active storage load is generated by the requirement to maintain interim outcomes for use in the next calculation, after which the previous answer can be discarded, freeing resources to update bindings and interim outcomes to facilitate progress to the next addend. This is central to an attentional control conceptualisation of *WM* (Kane et al., 2001). While the ABC variable-value mappings introduced in the retention condition appear to contribute to *WM* demands (by taking up bindings), they do not require direct access at any point until after the equation. Thus, these can be relegated to a long-term store during the processing. Meanwhile, the access condition introduces an additional step to solving the equation, in that the XYZ mappings are linked to the previously presented ABC mappings (i.e., rather than XYZ=143 as in the control and retention mappings, the access mappings are XYZ=ABC=143), requiring constant direct access to the ABC mappings. In the RC framework, this results in embedded relations that cannot be easily chunked without an appropriate strategy (Halford et al., 1998) and which must be kept directly accessible throughout the problem. In a fixed-access condition, where the XYZ variables are mapped in a consistent, linear order to ABC, there is *systematicity*. The constant, directly intuitive mappings between the ABC bindings (e.g., A=1, B=4, C=3) are systematically mapped to the fixed-order XYZ bindings (e.g., X=A, Y=B, Z=C). Systematicity exploits this common, natural ordering of fixed mappings, resulting in strategic reduction of the load on *WM* (e.g., ABC can be represented as the simple order 143, and XYZ can be directly mapped to this order as 143, also). The facilitation provided by this simplified representation manifests in a single, durably bound chunk. In a random-access condition, where the XYZ variables are mapped in a different alphabetical ordering (randomly) to ABC, systematicity breaks down. Each separate mapping (A=1, B=4, C=3) must be kept independent so they can be flexibly adapted to the random order of XYZ. Figure 4.1 represents these conditions schematically, demonstrating how only random-access requires a higher capacity for relational integration.

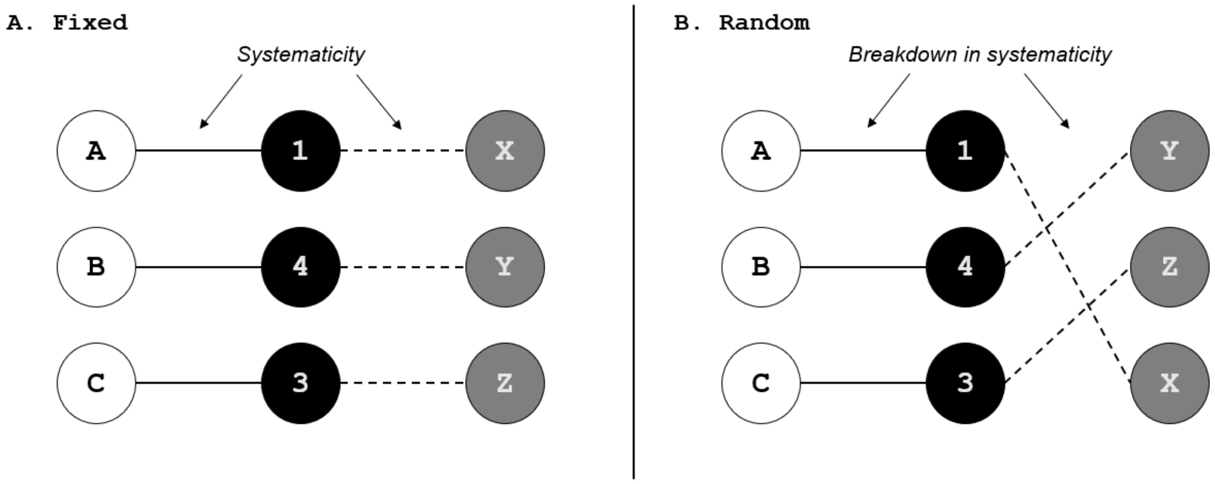


Figure 4.1. Schematic representation of the mappings required in fixed-access vs. random-access. In fixed-access (A), the constant order of XYZ results in systematicity, because it matches directly to the ordering of ABC. This systematicity can be exploited so that only one binding is kept active during the arithmetic (XYZ=143). In random-access (B), the systematicity of the constant XYZ order breaks down: the ABC order cannot be chunked down because participants do not yet know which order XYZ will appear in. For random-access, participants have two options: either maintain three separate bindings [A(1,X), B(4,Y), C(3,Z)] or rearrange the bindings into the updated YZX ordered chunk of 431. In either option, relative to fixed-access, there is an additional cognitive load generated which we argue is the binding load that those with a higher capacity for relational integration will be better able to manage.

In sum, the psychological implication for random-access is that an additional source of uncertainty in mappings is introduced. The convenience of naturally ordered, fixed mappings cannot be applied to random mappings, forcing a need for multiple, separate bindings. In related work, Chuderski (2014) found that the largest contributor to task performance on a relation monitoring task was the number of bindings that needed to be established simultaneously. In the monitoring task, participants have to monitor a 3x3 grid of strings and respond whenever a match (e.g., across a horizontal or vertical line) occurred that corresponded to a pre-determined match rule. Chuderski found that when the match rule necessitated independent bindings (e.g., when all matching strings had to be different) which could not be systematically chunked, performance was drastically impaired as the number of strings involved in the match increased; while match rules that could be systematically chunked (e.g., when all matching strings had to be the same) did not suffer this same cumulative penalty as the number of strings increased. In a similar way, we expect the largest detriment to performance under random-access condition, as it requires three independent bindings, rather than one systematically chunked binding.

Finally, we consider the relationship between WM and Gf. Commonly, WM has been viewed as a subordinate system to the more ethereal and (purportedly) immutable Gf, with measures of WM used to predict (Ackerman et al., 2005) or train (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008) performance on Gf tasks. However, there is a growing consensus that similar limiting factors act on both WM tasks and Gf tasks (Oberauer et al., 2007; Shipstead et al., 2016). Although Gf tasks typically involve additional complexities and demands, a limit in the number of bindings that can be established causes similar detriments to performance in both WM and Gf tasks. A binding explanation of the complex-span was provided earlier, but consider the quintessential example of a Gf task, Raven's Advanced Progressive Matrices (APM). The APM initially involves inducing complex relations among patterns. Individual elements are complex, as each cell is composed of multiple features such as lines or shapes. Lines may *inter alia* be straight, wavy, dotted, or differ in orientation; shapes may *inter alia* differ in size, shading, numerosity, or form. The specific rules that dictate how APM element features are related (or not) within a cell, or across rows or columns, must be induced by the participant (Verguts & De Boeck, 2002), then applied to the response options to derive the solution. At the simplest level, APM entails induction and then application of multiple unknown but discoverable rules about relations. However, *WM* is involved to the extent that these rules can be represented as a relational structure within the direct access region. Insufficient binding capacity limits the problem space available. Where rules can be represented as a schema, the integration of features (e.g., line or shape features) with rules is then required to generate the corresponding end-piece of the pattern (Oberauer et al., 2008). Thus, Gf tasks overlap with simple WM tasks to the extent that they both involve establishing temporary bindings. However, the more complex Gf tasks also require an initial induction of rules.

Oberauer et al. (2008) put forth considerable evidence for a binding conceptualization of the WM-Gf overlap, finding that relational integration tasks could explain variance in Gf over-and-above that accounted for by more traditional store-and-process WM tasks. Further, this predictive ability was not contingent on whether there were storage demands involved in the relational integration tasks, implicating the establishment of bindings or the construction of the relation as more important to the relationship than merely storing information over time. Similar outcomes were found in Chapter III, where the Latin Square Task (a matrix-style relational integration task with active storage components) performed better as a predictor of Gf when storage demands in the task were stripped by allowing participants to dynamically fill interim cells of the matrix. These results indicate the Latin Square Task can function as a more complex WM task (e.g., Birney et al., 2012) but that it may function better when the active storage demands are minimized and the task primarily measures relational integration. Additional evidence was contributed by Chuderski (2014), as the relation monitoring task (described earlier) also predicted Gf over-and-above other *WM* tasks, despite having no explicit storage requirements. Interestingly, Chuderski also found that, despite task performance being dictated by the number of bindings, the number of bindings did not differentially predict variance in Gf (as measured by APM). A similar result was concluded in Chapter III, where the LST seemed to provide qualitative differences at the 2D and 4D level, rather than a linear difference including the 3D level. Despite these studies agreeing that relational integration is important in both WM and Gf tasks, it does not seem that increasing the number of bindings necessarily increases the link between the two constructs. Conversely, the current study’s manipulation of random-access against fixed-access in the

ACT would directly contrast three item-specific bindings to one generic binding (natural-order).

The studies described here indicate that the WM and Gf tasks share common resources in binding capacity. In the current experiment, we aim to contribute to understanding on this common mechanism of binding by demonstrating how manipulations of bindings in the ACT impact on task performance and on the relationship to classic measures of WM (complex-span) and Gf (APM). The covariability between the complex-span and APM should represent the common resources shared by WM and Gf, which research (Chuderski, 2014; Oberauer et al., 2008) indicates is binding and relational integration.

The hypotheses for this experiment are associated with explaining how variability in ACT item response differentially demands resources common to the binding in WM and Gf. If different variants of the ACT demand this common resource differently, then we expect different patterns of associations. Hypotheses 1a and 1b are replication tests of Oberauer et al. (2001) while hypothesis 1c is a test of the additional binding demand incurred through random access as opposed to fixed access. The second hypothesis set is based on our task analysis of the ACT and moderation predictions focused on testing the common functions of WM and Gf.

Hypothesis 1: Identifying task demands (Costs)

The first hypothesis set compares the difficulty of the conditions to identify which manipulation produces the greatest demand and thus, has the greatest performance cost associated with it. Following from Oberauer et al.'s (2001) finding that a storage load only impacted processing performance when it had to be accessed during the processing, it was hypothesized (H1a) that retention items will incur no cost compared to (i.e., be no different from) control items because the ABC storage load can be relegated to long-term storage during the processing. Also following Oberauer et al., it was hypothesized (H1b) that access items will incur a cost over (i.e., be more difficult than) control and retention items, as the ABC storage load must be kept active in the direct-access region during the processing, as opposed to being relegated to a long-term store. Finally, for our novel manipulation of random-access vs. fixed-access, it was hypothesized (H1c) that random-access will incur a cost over (i.e., be more difficult than) fixed-access items, as the random administration necessitates three independent mappings that cannot be systematically chunked. This follows the results of Chuderski (2014), who found additional bindings greatly increased task demands.

Hypothesis 2: Predicting WM and Gf through Costs

The second set of hypotheses are concerned with identifying the functional source of the aforementioned costs. Each cost described has been conceptualised and operationalized to specifically demand secondary storage (outside the direct access region; *retention-cost*), direct access (*access-cost*), and independent binding costs (*binding-cost*). Each of these costs have been considered common to both WM and Gf (Shipstead et al., 2016) and layering these costs within a single task allows us to isolate the contributing function. We operationalise the WM-Gf commonality as a binding factor representing the variance shared by APM and a complex-span task. We predict that individuals' performance on this factor will moderate costed performance (i.e., retention-cost (H2a), access-cost (H2b), and binding-cost (H2c)) to the extent that the cost is common to WM and Gf.

4.2. Method

4.2.1. Participants

This study is structured as a multi-level, within-person design with binary responses. Intended sample-size was estimated using Optimal Design (Raudenbush et al., 2011) for power = .80. Effect size estimates were taken from Experiment 1 (Figure 2, *n* = 36) of

Oberauer et al. (2001). Mean proportion correct for the control and retention conditions (≈ .95). The access items were slightly more difficult (≈ .78). The ‘reasonable range’ was estimated from these values. While each person responded to 24 items in total, cluster size for estimation was set at the more conservative range of 6 items (smallest interaction and main effect cluster size). With these parameter values, estimated sample size for H1 on the main effects (costs) was 60. We first attempted to meet this recommendation with surplus, collecting 64 participants. While we were generally confident with the capacity for a robust experimental test of H1 with this sample size, there is little data against which to reliably estimate the moderation effect-size for H2, and thus we recognise the possibility of a type 2 error (not finding significant differences where differences exist). We attempted to account for this possibility by collecting a further 58 participants, bringing the total to 122, a more typical sample size for individual differences research (Detterman, 1989; Marszalek, Barber, Kohlhart, & Holmes, 2011). These two sets of participants were collected in separate batches. Because we conducted the analysis on the first batch before collecting the second batch, this method of split-batch data collection raises the risk of a type 1 error. In line with the recommendations of Simmons, Nelson, and Simonsohn (2011), we outline our method of recruitment and the results of these first analyses to be fully transparent on the potential risk of our split-batch recruitment. Our method of recruitment allows for approximately 65 participants to be recruited per research period. Thus, the intended sample size was not decided arbitrarily but with the intention of collecting maximum participants within two research periods. When accounting for some potential loss of sample, this aligned with our H1 parameter estimates (60) and with our intended sample for H2 (100+). The results of the first analysis (*n* = 64), which follow the same pattern of results as those presented here, are included in Appendix B.

The participants were 128 first-year psychology students at the University of Sydney who participated in exchange for course credit. Participants had 90 minutes to undertake the three key tasks (and two additional tasks not reported here). Three participants were removed from the analyses as they did not attempt one or more of the key tasks (ACT, SSPAN, or APM) in the time provided. A further three participants were removed for scoring unreasonably low (< 3 SD of mean) on at least one task while also having unreasonably short total time taken (< 1 SD mean) on that task, indicating they were not properly engaged in the task. Finally, 6 participants who scored less than 80% correct on the processing component of the SSPAN task, indicating they did not follow task instructions, were excluded. Because most analyses incorporated the SSPAN, for simplicity, we simply excluded these participants from all analyses also. This resulted in a final sample size of 116 (90 females) with a mean age of 20.06 (SD = 4.09) years.

4.2.2. Measures

Arithmetic Chain Task (ACT)

Participants completed four ACT blocks of problems (one for each condition: control, retention, fixed-access, and random-access), presented in a random order. Participants received instructions on each of the problem types (the same set of instructions were given for access-fixed and access-random) and received reminders before each block on these types. Trials were randomly generated such that all displayed digits were between 1 and 7, and intermediary and final answers were between -9 and +9. Participants were informed of these restrictions. There were six items generated for each block, totalling 24 unique items. Two scores were derived from the ACT for each item; accuracy (0-1) and, for items with a recall component, recall (0-3). Responses were self-paced, but the program terminated after a total of 30 minutes.

Control items were standard problems that entailed mentally substituting variable-value relations (e.g., X=2, Y=1, Z=4) provided in the top half of the screen into equations where each operand was displayed one-at-a-time at a pace controlled by participants (see Figure 4.2). Participants continued to press the spacebar until all 7 operands had been displayed, at which point a textbox would appear allowing the participant to type in their answer. Participants had to incorporate the variable-value mappings displayed above the equation in order to derive the solution. The variables (X, Y, and Z) were integrated into the arithmetic chain at random. Correct answer feedback was displayed for two seconds, before moving on to the next item. Unlike Oberauer et al. (2001), there was only one mode of presentation: all operations stayed on the screen until the equation was solved.

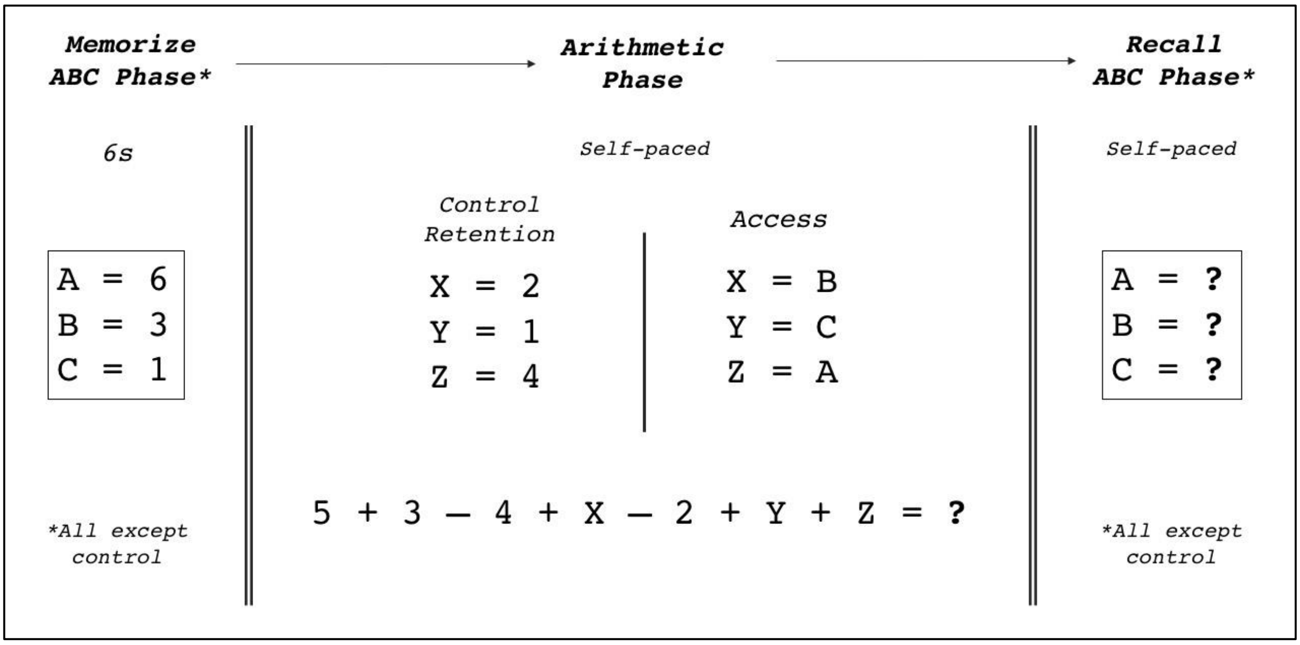


Figure 4.2. Graphical representation of an item in the task. From left-to-right: in all conditions except control, participants memorize the ABC mappings for 6s. Participants then solve the arithmetic which incorporates the XYZ mappings. In the control and retention conditions, the XYZ correspond directly to numbers. In the access conditions, the XYZ correspond to ABC (either in fixed order or, as shown here, random order). Finally, in all conditions except control, participants recall the ABC mappings.

Retention items were identical to control items, with the exception that the participants were given 6 seconds prior to engaging in the arithmetic to memorise three variable-value mappings (e.g., A=6, B=3, C=1) to be recalled at the end of the trial. Feedback was given after each recalled response.

Fixed-access items were similar to retention items, except that the *XYZ* variable-value mappings were directly and consistently linked to the *ABC* variable-value mappings (e.g., A=6, B=3, C=1; and always, X=A, Y=B, Z=C). Again, participants were asked to reproduce the digits corresponding to *ABC* after the equation had been solved. Thus, unlike the retention condition, the ABC mappings were required for the mental arithmetic.

Random-access items were similar to fixed-access items, so much so that they did not have a separate set of instructions or notifications. The only difference to fixed-access was that the *XYZ* variable-value mappings were directly but *randomly* linked to the *ABC* variable-value mappings (e.g., A=6, B=3, C=1; and say, X=B, Y=C, Z=A). As for the other conditions, participants were asked to reproduce the digits corresponding to *ABC* after the equation had been solved.

Symmetry Complex-Span (SymSpan)

The same SymSpan reported in Experiment 1 of Chapter III was used here, based on the complex span task reported by Kane et al. (2004). Consistent with the CSPAN paradigm, the processing component entailed judgments (yes/no) of whether a displayed pattern was symmetrical along the vertical axis. The storage component was a spatial-memory updating task in which the location of a red square randomly presented in a 4x4 grid for 850ms needed to be remembered and recalled at the end of the set. Set sizes varied between two and five squares. The score analysed was the total number of correctly recalled squares across the task with a theoretical range of 0 – 28. This ‘total squares’ partial scoring method was preferred to pick up additional variance that would be otherwise discarded by only considering fully correct sets (Redick et al., 2012). As described earlier, participants needed to score at least 80% on the processing (symmetry) component to be included in the analyses.

Raven's Advanced Progressive Matrices

Gf was measured using the abbreviated 20-item version (odd items + items 34 and 36) of set II of the APM (J. C. Raven, 1941), as in Chapter III. Participants had 20 minutes to complete as many of the 20 items as possible.

Relational Binding (RB)

A factor analysis (with principal factor extraction) of the APM and SSPAN scores was conducted using the R package 'psych::fa' (Revelle, 2018) to derive common-factor scores to represent what is common to both *WM* and *Gf* in these prototypical measures of these constructs (Shipstead et al., 2016). We refer to this as *RB*, a relational binding factor. Empirically, this represents the intersection between the tasks (Shipstead et al., 2016); conceptually, this represents a participants' capacity for binding (Oberauer, 2009a). Although we have argued extensively for our theoretical position (that this intersection represents relational binding), there are of course other interpretations of this *WM-Gf* overlap (e.g., Kane et al., 2001) that may be applicable. For those interpretations, our reference of this *RB factor* can serve simply as a label for 'the overlap between *WM (SSPAN)* and *Gf (APM)'*. The eigenvalue of the first component accounted for 60.48% of the total variance, and 20.90% of the extracted shared variance. The variable was then the extracted factor scores using regression.

4.3. Results

There was a small number of ACT item-level data missing at random (2.6%) from incomplete timeouts. In total, the data analyses were based on 2744 observations from 116 participants (a complete data set would have provided 24 items x 116 participants = 2784 observations). All analyses were performed with R version 3.5.2 (R Core Team, 2018). Plots were produced with the 'sjPlot' (Lüdecke, 2017) and 'ggplot2' (Wickham, 2009) packages. Hypotheses were tested by modelling item responses using a mixed-effects logistic regression approach as implemented in the ‘glmer’ procedure from ‘lme4’ (1.1.17) package (Bates, Maechler, Bolker, & Walker, 2017). Effects of each condition was conceptualized as costs (demands on WM) and costs were operationalized using contrasts as the decrease in log-odds chance of getting an item from a given test condition (e.g., retention) correct compared to the chance of getting an item from the reference condition (e.g., control) correct. The relationship with RB was then considered by modelling RB as an interaction term with each contrast to determine the extent to which RB moderated the influence of each cost. A random-intercept model with condition as both a fixed and random effect (Model 1) was used to derive estimate of the overall descriptive statistics for each condition as reported in Table 4.1 and the relationship with RB as plotted in Figure 4.4. Model 2 tested orthogonal effect-contrasts consistent with the stated hypotheses (contrast coefficients are provided in Table 4.2 alongside the main analyses). We first analyse recall accuracy for the ABC variables.

4.3.1. Recall accuracy

Recall of the ABC variable-value mappings was above 85% in all conditions, the cut-off typically used for the secondary complex-span task. However, a linear-mixed effects regression analyses of mean recall accuracy on condition dummy coded (with subjects as level 2) revealed expected differences. Recall of the ABC mappings on retention items was significantly poorer than both types of access items (Access-F: $\beta = 0.27$, $CI_{95\%} = [0.20, 0.35]$, $p < .001$; Access-R, $\beta = 0.19$, $CI_{95\%} = [0.11, 0.26]$, $p < .001$), and the two types of access items differed in recall accuracy ($\beta = 0.08$, $CI_{95\%}$ [0.01, 0.16], $p = .030$). As seen in Figure 4.3, this produced a different pattern of scores depending on whether recall accuracy was or was not factored into the scoring of the arithmetic, with retention seeing a notable drop in performance when making arithmetic performance for each item conditional on also perfectly recalling all three ABC mappings for that item (we refer to this as ‘absolute’ scoring, as in the item was absolutely correct in both arithmetic and recall). A series of paired *t*-tests revealed all three experimental conditions suffered a detriment in accuracy (at $p < .001$) as a result of making accuracy conditional on both correctly answering the arithmetic and perfectly recalling the three ABC variables, with the largest effect in the retention condition (retention $d = 1.03$; Access-fixed $d = .38$; Access-random $d = .47$). Given these differences, it was worth discussion on which scoring method was preferred (arithmetic-only or absolute). Although generally only a minimum threshold of performance is needed (e.g., 85%+) for secondary tasks in complex-span paradigms, there was a substantial theoretical reason to prefer ‘absolute’ scoring in the ACT. By only looking at trials where recall performance was perfect, we could safely assume that all the correct trials have satisfied the basic maintenance component of the task (i.e., the ABC variables were successfully maintained and recalled), so any differences that emerge between conditions (e.g., access-fixed versus access-random) must be due to demands associated with the arithmetic. Given that our hypotheses are based on the unique, incremental demands of each condition, targeting the trials where the arithmetic response was incorrect *despite* correctly recalling the maintained variables would help to isolate the experimental effect of each condition. For instance, it is possible that a participant may incorrectly answer an access-random item either due to the binding cost impacting on the arithmetic *or* because they lost the ABC variables. Losing the ABC variables (e.g., through decay) gives the participant no chance of solving the item correctly, even if they are perfectly capable of handling the binding burden of access-random. Although the high recall performance (see Table 4.1) means these trials are not particularly common (i.e., the trials where a participant loses the ABC variables through reasons unrelated to the condition’s cost), using absolute scoring nonetheless provides a purer measure of the cost. This scoring method was also consistent with Oberauer et al. (2001), who used absolute scoring as standard for the analyses.

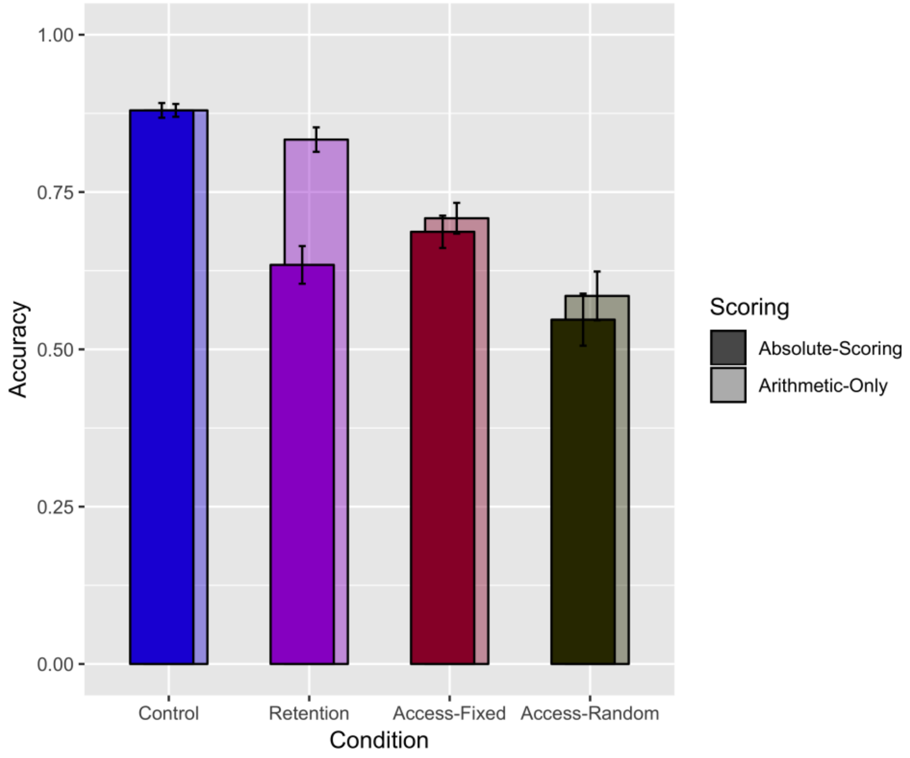


Figure 4.3. Plot of accuracy across four ACT conditions. The overlayed graph represents the difference when performance is also conditional on correct recall. Arithmetic-only scoring (light bars) is the proportion of trials where the arithmetic was correct disregarding recall performance, while absolute scoring (solid bars) is the proportion of trials where both the arithmetic and recall was completely correct.

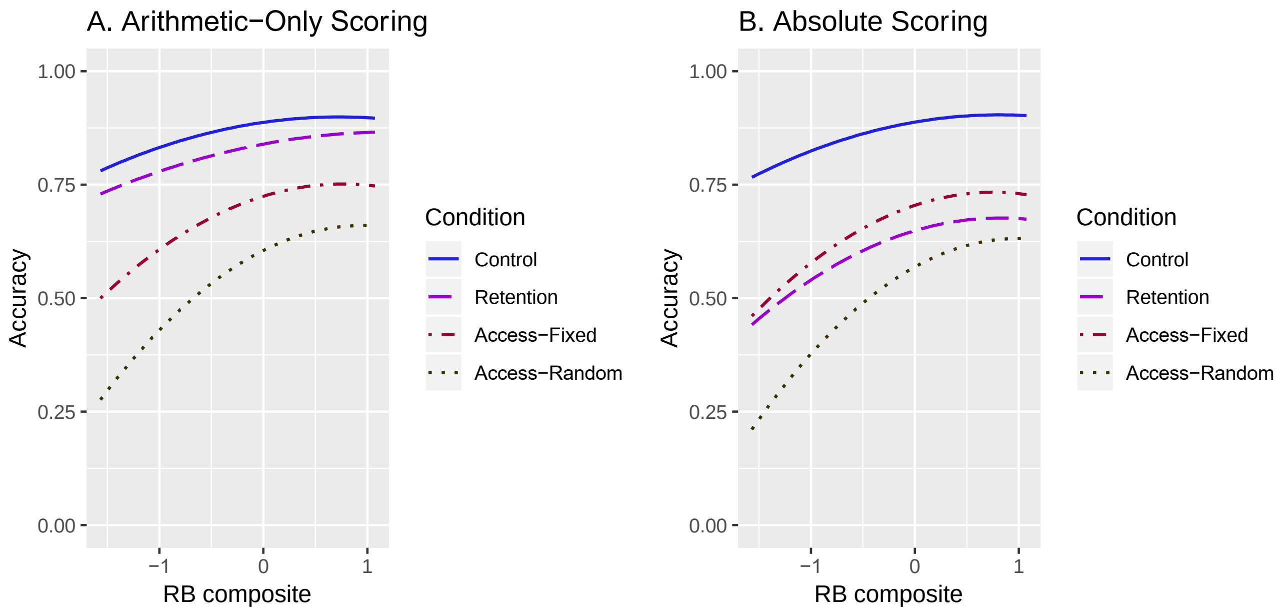


Figure 4.4. Relationship of each ACT condition to RB composite, split by (A) arithmetic-only scoring and (B) absolute scoring.

Table 4.1. *ACT Descriptive Statistics (with absolute scoring for accuracy)*

		Accuracy		Recall	
	N	Mean	SD	Mean	SD
Control	114	0.88	0.06	-	-
Retention	113	0.63	0.16	0.84	0.15
Access-F	116	0.69	0.14	0.94	0.09
Access-R	115	0.55	0.22	0.91	0.11
SSPAN	116	0.75	0.16		
APM	116	0.61	0.19		

r(APM,SSPAN) = .22, *p* < .05

4.3.2. ACT accuracy and RB demands

As described prior, the following analyses use absolute scoring for the ACT. As seen in Figure 4.3, trends for item difficulty were generally in the theoretically expected direction. Control items were significantly easier than all other items on average ($CI_{95\%}$ = [-1.832, -1.289], p < .001). Contrary to H1a, there was evidence for the presence of retention-costs (performance decline on retention items relative to control items), $CI_{95\%}$ = [-1.821, -1.201], p < .001, though this cost was marginally not significant when disregarding recall performance, $CI_{95\%}$ = [-0.705, 0.017], p = .062. Figure 4.5A demonstrates that individual differences existed in these retention-costs though, contrary to H2a, these differences were not determined by RB ($CI_{95\%}$ = [-0.798, 0.205], p = .247), as seen in Table 4.2. This suggests that the additional load of retention over and above standard ACT items that a participant may experience, is not a simple function of relational binding capacity.

Contrary to H1b, there was no evidence for an access cost, as performance on access items overall was similar to retention items ($CI_{95\%}$ = [-0.318, 0.171], p = .553; Table 4.2). As with retention-costs, although there were individual differences (Figure 4.5B), they were not determined by *RB* ($CI_{95\%}$ = [-0.064, 0.768], p = .097), thus failing to support H2b.

Consistent with expectations of H1c, there was evidence for a binding-cost. Performance on access-random items was significantly poorer than that of access-fixed items

(CI$_{95\%}$ = [-0.930, -0.391, p < .001; Table 4.2). However, unlike retention-costs and access-costs, and consistent with H2c, there does appear to be evidence that the individual differences in binding-costs (Figure 4.5C) can be understood to some extent by demands on *RB* capacity (CI$_{95\%}$ = [0.102, 1.043], p = .017).

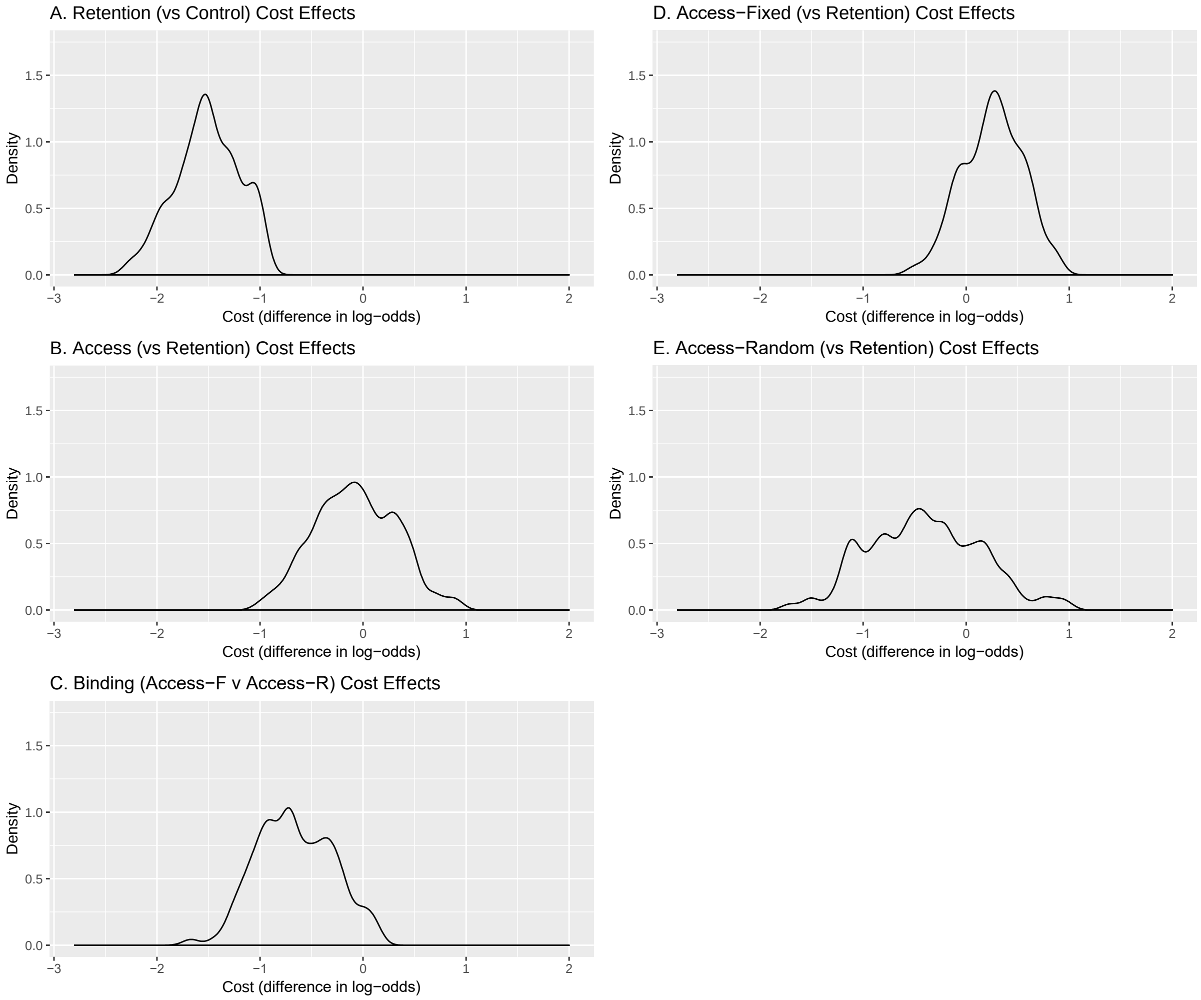


Figure 4.5. Density distributions of individual differences for (A) retention costs, (B) access costs, and (C) binding costs. Access costs (relative to retention) are decomposed to (D) access-fixed only and (E) access-random only.

Table 4.2. *Fixed and Random Effects Estimates of Planned Contrasts (ACT absolute scoring)*

		Fixed Effects				**Random Effects**	**Contrast Coding**			
Predictors	**Model**	**Log-odds**	**se**	**CI**	**p**	**tau**	**Control**	**Retention**	**Access-Fixed**	**Access-Random**
(Intercept)	1	0.946	0.088	0.774, 1.119	**<0.001**	0.62	-	-	-	-
Retention (vs Control) Cost	1	-1.511	0.158	-1.821, -1.201	**<0.001**	0.25	-1/2	1/2	0	0
Access (vs Retention) Cost	2	-0.074	0.125	-0.318, 0.171	0.553	0.52	0	-2/3	1/3	1/3
Binding (Fixed vs Random) Cost	1	-0.661	0.138	-0.930, -0.391	**<0.001**	0.46	0	0	-1/2	1/2
RB moderator	1	0.575	0.149	0.283, 0.868	**<0.001**	-	-	-	-	-
Retention (vs Control) Cost x RB	1	-0.296	0.256	-0.798, 0.205	0.247	-	-	-	-	-
Access (vs Retention) Cost x RB	2	0.352	0.212	-0.064, 0.768	0.097	-	-	-	-	-
Binding (Fixed vs Random) Cost x RB	1	0.572	0.240	0.102, 1.043	**0.017**	-	-	-	-	-
Fixed (vs Retention) Cost	3	0.256	0.134	-0.007, 0.520	0.056	0.38	0	-1/2	1/2	0
Fixed (vs Retention) Cost x RB	3	0.066	0.225	-0.376, 0.508	0.771	-	-	-	-	-
Random (vs Retention) Cost	4	-0.404	0.150	-0.699, -0.110	**0.007**	0.89	0	-1/2	0	1/2
Random (vs Retention) Cost x RB	4	0.638	0.261	0.127, 1.149	**0.014**	-	-	-	-	-
Control vs Rest	2	-1.561	0.138	-1.832, -1.289	**<0.001**	0.13	-3/4	1/4	1/4	1/4
Access vs Rest	1	-0.830	0.111	-1.046, -0.613	**<0.001**	0.32	-1/2	-1/2	1/2	1/2
Control vs Rest x RB	2	-0.062	0.222	-0.497, 0.373	0.781	-	-	-	-	-
Access vs Rest x RB	1	0.204	0.185	-0.159, 0.567	0.272	-	-	-	-	-
orthog1	*3*	-	-	-	-	-	*-2/3*	*1/3*	*1/3*	*0*
orthog2	*3*	-	-	-	-	-	*1/4*	*1/4*	*1/4*	*-3/4*
orthog3	*4*	-	-	-	-	-	*-2/3*	*1/3*	*0*	*1/3*
orthog4	*4*	-	-	-	-	-	*-1/4*	*-1/4*	*3/4*	*-1/4*
N = 2,744 observations; Conditional R^2 = .304						**σ^2= 3.29**				

Notes: To test the contrast of interest, two sets of orthogonal contrasts were needed. The column Model indicates which model the estimates have come from and the notation below specifies the full model (orthog1-4 were needed to ensure orthogonality). Binding was tested in models 1 and 2 and as expected, produced identical estimates for all effects in both models. The follow-up effects of access-fixed vs retention and access-random vs retention were tested in Models 3 and 4. Models tested are as follow:

Model 1: glmer(ACT ~ 1 + bindingC*RB + retentionC*RB + AccessVrestC*RB + (1 + bindingC + retentionC + AccessVrestC | subject)
Model 2: glmer(ACT ~ 1 + bindingC*RB + accessC*RB + ControlVrestC*RB + (1 + bindingC + accessC + ControlVrestC | subject)
Model 3: glmer(ACT ~ 1 + accessFC*RB + orthog1*RB + orthog2*RB + (1 + accessFC + orthog1 + orthog2 | subject)
Model 4: glmer(ACT ~ 1 + accessRC*RB + orthog3*RB + orthog4*RB + (1 + accessFC + orthog3 + orthog4 | subject)

Thus, of the three theoretical loads investigated, there was evidence for retention-costs and binding-costs on accuracy.[9] However, only binding-costs were associated with individual differences in our composite *RB* factor, which was defined as what is common to the APM and symmetry CSPAN tasks. Table 4.3 displays the simple correlation matrix, which makes the pattern between *RB,* the constituent *RB* measures (APM and SSPAN), and the ACT conditions clear: retention correlates with SSPAN ($r = .24$, $p = .010$) but not APM ($r = .02$, $p = .867$); while access-fixed correlates with APM ($r = .28$, $p = .003$) but not SSPAN ($r = .06$, $p = .531$). Access-random however, is the only ACT condition that correlates with both SSPAN ($r = .23$, $p = .014$) and APM ($r = .35$, $p < .001$). A linear regression with the four ACT conditions predicting *RB* revealed that access-random was the only condition to predict a significant, unique proportion of the variance in *RB* ($sr^2 = .069$, $p = .004$),[10] solely predicting almost 40% of the total variance accounted for in *RB* ($R^2 = .174$).

Table 4.3. *Simple correlations between ACT conditions and criterion tasks.*

	Control	Retention	Access-F	Access-R	SSPAN	APM
Control	**.25**					
Retention	.31**	**.52**				
Access-F	.33**	.38**	**.51**			
Access-R	.33**	.36**	.43**	**.68**		
SSPAN	.13	.24**	.06	.23*	-	
APM	.34**	.02	.28**	.35**	.21*	-
RB factor	.30**	.16*	.22*	.37**	*.78***	*.78***

Notes: * p < .05; ** p < .01; **bold** on diagonal represents reliability (Cronbach's α)

[9] As with Oberauer et al. (2001), we also performed the analyses using response time (RT) rather than accuracy (log-transformed to adjust for long RT outliers). The pattern of results remained similar, with a clear linear increase in time taken similar to the arithmetic-only scores seen in Figure 4.3 (control M = 4.22s, SD = .12; retention M = 4.31s, SD = .14; access-fixed M = 4.43s, SD = 1.34; access-random M = 4.51s, SD = 1.45). This meant that, when using log-RT, there was evidence for retention-costs ($CI_{95\%} = [0.11, 0.17]$, p < .001), access-costs ($CI_{95\%} = [0.27, 0.32]$, p < .001), and binding-costs ($CI_{95\%} = [0.07, 0.10]$, p < .001). The presence of retention-costs was contrary to Oberauer et al.'s results, though may be related to the exclusion of manipulations on item-load and participant age. In an identical pattern to the accuracy results, *RB* appeared to determine the variability between participants seen in binding-costs ($CI_{95\%} = [0.11, 0.17]$, p < .001), but not in retention-costs ($CI_{95\%} = [-0.00, 0.05]$, p = .109) or access-costs ($CI_{95\%} = [-0.02, 0.03]$, p = .910). Although interesting, we have no specific hypotheses related to RT and the results are similar enough to those seen in the accuracy data, so we do not mention RT further.

[10] Control $sr^2 = .027$, $p = .065$; Retention $sr^2 < .001$, $p = .890$; Access-Fixed $sr^2 < .001$, $p = .708$.

4.4. Discussion

Rather than working memory capacity acting largely as a distinct subordinate function of fluid intelligence, there is an emerging consensus that the *WM-Gf* link (e.g., Ackerman et al., 2005; Engle, Tuholiski, et al., 1999; Kyllonen & Christal, 1990; Unsworth & Engle, 2005) can be understood as the outcome of common functions dictated by the strength and flexibility of relational bindings between integrated representations (Chuderski, 2014; Oberauer et al., 2007; Shipstead et al., 2016). In the current study, we manipulated a single task (the Arithmetic Chain Task; Oberauer et al., 2001) in order to differentially demand retention, access, and binding. The layering of manipulations allowed us to pinpoint the contributing functions while reducing task artefacts typically present in comparing multiple task formats. Our manipulations began on a similar premise to Oberauer et al. (2001), in distinguishing retention from access by comparing how stored contents (the ABC mappings) were assessed during arithmetic performance: either passively (recalled at the end of the task only) or accessed during active processing (incorporating the mappings as part of the arithmetic). We extend on Oberauer et al.'s (2001) research by comparing fixed-access to random-access which prevents systematic chunking (Halford et al., 1998). Our findings partially replicated Oberauer et al. by finding that access does incur a larger cost to arithmetic processing over passive retention – but we have demonstrated this only occurs if the stored variable mappings must be accessed in a random order. This indicates that Oberauer et al. would not have found their results (that access incurred a greater load than retention) if they had made the seemingly minor change of not randomly ordering the XYZ=ABC mappings. This is demonstrated most clearly in Figure 4.5, where breaking down the access cost (Figure 4.5B) into its constituent access conditions (Figures 4.5D and 4.5E) produced markedly different outcomes: only random-access shows a significant cost over retention (the distribution is centred below 0). Critically, we also discovered that performance on random-access specifically (i.e., layered over an otherwise identical task format in fixed-access) is uniquely influenced by what is common to WM and Gf – which we have argued is the demands of binding (Chuderski, 2014; Oberauer et al., 2008).

Figure 4.6 adapts Oberauer's (2009) concentric model, including the source of demands specific to each version of the ACT presented in the current study. *Retention costs* were defined as the demand imposed by encoding and maintaining additional task-irrelevant mappings for later recall. These mappings are theorized to be stored in long-term memory (outside the region of direct access) for the duration of the task. Although we identified a retention cost that Oberauer et al. (2001) did not, this cost was largely driven through a failure to recall these task-irrelevant mappings, as opposed to a load influencing the arithmetic itself (see Figure 4.3). In contrast, *access costs* incorporated task-critical mappings, requiring establishing and maintaining bindings within the direct access region of *WM* throughout the task. We introduced an additional cost associated with ensuring multiple, flexible bindings in the direct access region by restricting systematicity through random rather than fixed mappings. Our proposition was that the systematicity that facilitates a single strong schema set where mappings are in a fixed serial order cannot be exploited when mappings are random, necessitating maintaining access to multiple independent bindings (Chuderski, 2014). The breakdown in systematicity results in unstable bindings that must be flexibly bound and unbound in light of the updated ordering only indicated during the executive processing of the primary arithmetic.

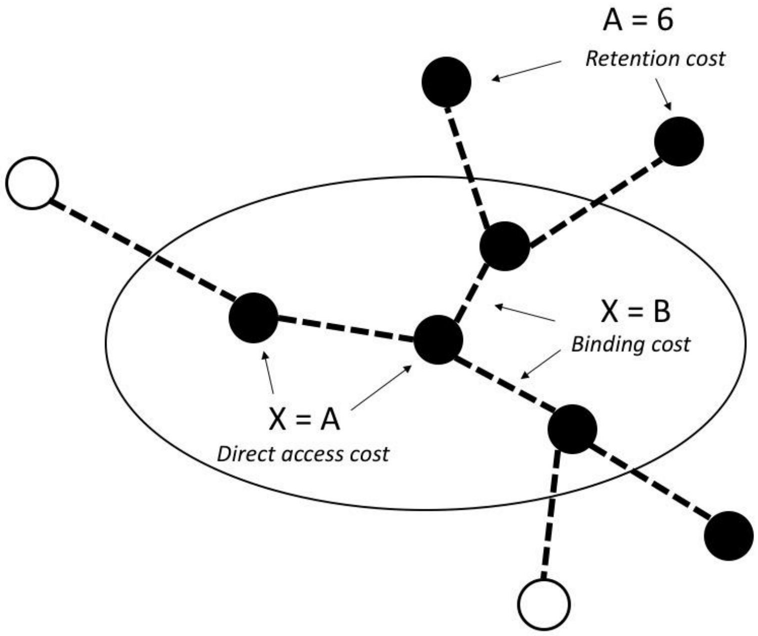


Figure 4.6. Diagram of Oberauer's (2009) concentric model of WM adapted to specify task manipulations in the featured Arithmetic Chain Task. Each small circle depicts a representation within memory, which can either be active (filled circle) or inactive (unfilled circle). The larger oval represents the *region of direct access*, a capacity-limited store where representations are active above threshold and available for immediate binding and further processing. Representations in the direct-access region can be connected into a common schema set by binding them into a related context. In the ACT, *retention costs* involve passively storing ABC mappings (e.g., A = 6) outside the direct-access region during the arithmetic processing. *Access costs* are incurred by ABC mappings which must be kept active in the direct-access region for use in the arithmetic processing (called upon in cases like X = A). *Binding costs* are incurred by ABC mappings which must be flexibly unbound and rebound into an updated order during the arithmetic processing (cases such as X = B).

Based on work such as Oberauer et al. (2008) and Chuderski (2014), we defined a *relational binding (RB)* factor as what is common to *WM* and *Gf* (defined by the SSPAN and APM, respectively). While it is unusual to run a factor analysis on just two variables, we argued that this was more appropriate than including the SSPAN and APM separately in a regression analysis, where their respective regression coefficients would reflect unique contributions, and the common features would be obscured as shared variance without direct assessment. Thus, the EFA was used to a create a simple *RB* indicator from prototypical measures to approximate what is common to *WM* and *Gf*. Performance on the *RB* factor indicates the extent to which participants performed well on what is common to *WM* and *Gf* – theorized to be the capacity for flexible binding.

Retention costs did produce a significant impact on performance over the control condition when recall was considered as part of the scoring, yet these costs were not associated with the *RB* factor. In the retention condition, passive storage demands were incurred by encoding a set of task-external mappings at the beginning of the trial and recalling them at the end of the trial. In this way, the unrelated storage could be relegated to long-term memory outside the direct-access region. The current results indicate that this passive retention is not associated with the *RB* factor. This manipulation of retention is different from traditional CSPANs, where repeated unrelated, trial-specific processing temporally overlaps with storage in which the running sequence of list items must be updated regularly. Since our retention involved encoding at the beginning and recall at the end, this storage was more passive than that required by the within-trial updating of CSPAN where the direct-access region is frequently probed with intermittent processing. Despite this, the CSPAN included in the current experiment correlated substantially better with retention than any of the other ACT conditions, while also providing a version of the ACT similar enough to the access conditions where the specific effect of access could be isolated.

In contrast to retention, *access* costs represent ABC mappings which must be kept active in the direct-access region during the arithmetic processing. The present results suggest that the direct-access region may be a source of capacity limits, but it is one that can be circumvented with systematic chunking of consistently-ordered bindings. We speculated that exploiting the fixed-order of mappings could systematically reduce the number of bindings from three to one. To account for this, we contrasted two access conditions: fixed and random. The conceptual difference between fixed and random is the flexibility of the bindings necessary to respond to the yet-unknown order. Although both access conditions put demands on the direct-access region, only the random-access requires maintaining three independent bindings. It is possible a systematic rearrangement of mappings could occur before the arithmetic processing has begun (but after the random order is revealed), but this still requires rapid binding and unbinding – a clearly isolated function above and beyond the otherwise identical fixed-access condition. There is a higher chance of losing the bindings during this rearrangement, and we observed small statistically significant differences in recall performance between the access conditions. While loss is a contributing factor, crucially, and consistent with Oberauer et al. (2007), Wilhelm, Hildebrandt, and Oberauer (2013), and our own task analyses incorporating systematicity (Halford et al., 1998), *binding costs* were significantly related to the *RB* factor, and this is by way of the random-access condition. Given that the fixed and random manipulations use an otherwise identical task format, this provides supporting evidence that what is common to WM and Gf is a capacity for flexible binding. It is worth reiterating the insights provided by this result. The binding costs inferred through the random-access condition already account for all other aspects of the ACT format. That is, the mental arithmetic involved in the core task and the additional burden of encoding, accessing, and recalling the ABC mappings through the task have already been accounted for. The exclusive component of random-access that remains after this incremental cost-analysis is the restriction that multiple bindings cannot be systematically reduced by way of fixed ordering. This restriction necessitates multiple bindings, and our results indicate that this is associated with the common factor between WM and Gf. As predicted by Oberauer et al.’s (2008) hypothesis, performance on WM tasks appears to be dictated by the binding capacity of the direct-access region. Here, we further demonstrate that the ability to rapidly establish and dissolve flexible bindings uniquely explains what is common to WM and Gf. In the current analyses, we labelled this commonality *RB* to represent our theoretical position. It is of course possible this commonality could be interpreted differently (e.g., controlled attention; Kane et al., 2001), though these interpretations would also need to provide a theoretical account of the difference between fixed-access and random-access, as we have done using systematicity (Halford et al., 1998).

4.4.1. Conclusion

In conclusion, our data suggest that it is not mere passive retention, nor systematic access, that defines the common WM-Gf link but rather, the ability to establish and maintain flexible bindings. In this way, CSPAN is a useful tool not simply because it taps storage capacity, but because the interim processing frequently interrupts the strength and stability of bindings of to-be-remembered elements. Passive retention of the to-be-remembered elements does not appear important, but providing direct access to durable, flexible bindings is. A version of the ACT which incorporates the temporal overlap of processing and storage (seen in CSPANs) between ABC mappings and the arithmetic may provide further insight into this notion, as may an *RB* factor defined through additional tasks. For now, our results provide preliminary but insightful evidence of the importance of a binding flexibility function in *WM*.

V. STUDY 3: THE RELATION MONITORING TASK

This chapter is published in Bateman, Thompson, and Birney (2019). [Bateman, J. E., Thompson, K., A., & Birney, D. P. (2019). Validating the relation monitoring task as a measure of relational integration and predictor of fluid intelligence. *Memory & Cognition, 47* (8), 1457-1468]. There are minor changes to terminology and flow to fit the thesis.

In the previous chapters, we found promising results for a relational approach to understanding WM and Gf. In Chapter III, the LST was found to perform even better as a predictor of Gf when active storage demands were minimized and instead, focus was placed on the relational demands. In Chapter IV, while the ACT involved either active or passive storage demands, the key manipulation of *access-fixed* against *access-random* demonstrated that increased binding demands associated with random ordering of to-be-remembered elements was crucial to linking the task to Gf. While these studies were mostly successful, they involved either a substantial manipulation of an established task (the LST) or an intricate manipulation of basic arithmetic (the ACT). In the current chapter, we turn towards a less well-known task (the Relation Monitoring Task) that has shown promise as a pure relational integration measure (Oberauer et al., 2008; Chuderski, 2014). The development of a relational integration task with inherently minimal storage demands and no superficially similar overlap with Gf matrix-style tasks is essential to establishing relational integration as a construct for measurement, more widely beyond simply understanding WM. Although the Relation Monitoring Task has shown success in prior studies as a measure of relational integration (Oberauer et al., 2008; Chuderski, 2014), the exact task specifications differ between studies. Thus, it is not yet known what is required in administering the task to make it a successful measure of relational integration. The aim of the current chapter is to validate the Relation Monitoring Task as a pure measure of relational integration and determine the task factors that contribute to its performance as a measure of relational integration.

5.1. Introduction to the Relation Monitoring Task

The *Relation Monitoring Task* (RMT) involves monitoring a grid (typically 3 x 3) of periodically changing stimuli (e.g., words or digits) and detecting relational matches that may appear across rows or columns, according to a pre-determined match rule (e.g., three numbers end in same last digit), before the array is updated with new stimuli. This simple task is hypothesized to load on a capacity for relational integration (Chuderski, 2014; Oberauer et al., 2008): the ability to connect multiple elements within working memory (WM). Relational integration is thought to be the cornerstone of higher-order intelligence (Chapter IV; Halford et al., 1998; Oberauer et al., 2008), required in well-established measures of fluid intelligence (Gf) (J. Raven, 1989), and forming the premise of analogical reasoning tasks (Sternberg, 1977). Indeed, the RMT has demonstrated a remarkable ability to predict performance on intelligence tasks (Chuderski, 2014; Krumm et al., 2009; Oberauer et al., 2008), despite involving no explicit (i.e., controlled) storage of information over time. This implies that the often-cited link between WM and Gf (Ackerman et al., 2005) may inadequately capture the important role of relational integration in Gf. However, because the RMT also involves rapid scanning, it is difficult to rule out theories of attentional control altogether (Engle & Kane, 2004). Although Chuderski’s (2014) experimental manipulations of visual interference have indicated that attentional control has minimal impact on RMT performance, these results are preliminary and the theoretical aspects of the task are still largely equivocal.

The purpose of the current report is to (a) replicate previous findings demonstrating the RMT predicting Gf over-and-above classic WM tasks; and (b) more comprehensively understand the factors that influence RMT performance and the relationship to Gf. To this end, three theoretically aligned RMT manipulations were developed and implemented. First, we varied the complexity of relations to be integrated, because the capacity to deal with complexity has been recognised as a core determinant of intellectual function (Birney &

Bowman, 2009; Stankov, 2000). Second, the amount of new information present in each trial was manipulated in an attempt to tease apart the role of visual scanning and attentional control in the RMT and its relationship with Gf. Finally, to explore the role of inhibition, we manipulated the amount of visual interference presented in each trial (Chuderski, 2014). In the following sections, we describe the background of the RMT and detail the rationale for these manipulations. Our experiment replicates prior research demonstrating the RMT's remarkable ability to predict Gf and reveals that attentional control does contribute – though is not imperative – to this ability. Instead, it appears that the core demand of the task – the ability to bind multiple elements into an integrated relation – is what is paramount to the relationship with Gf.

The RMT was originally featured in Oberauer, Süß, Wilhelm, and Wittman's (2003) analysis of WM. Participants are presented with a 3 × 3 array of three-digit strings (see Figure 5.1 in the Method). In the standard version of the task, participants are asked to validate whether there is a row or column in which a particular rule holds (e.g., all digit strings end in the *same* last digit). In Oberauer et al. (2008), a re-analysis of the data revealed strong correlations with latent constructs of intelligence, particularly Gf, typified by Raven's-style (J. Raven, 1989) abstract reasoning tasks. Buehner, Krumm, Ziegler, and Pluecken (2006) and Krumm et al. (2009) found similar correlations between the RMT and Gf. The strong overlap between the RMT and Gf is supportive of a theory we are terming the *relational integration hypothesis:* the theory that performance in Gf is most fundamentally and ultimately limited by the capacity for relational integration, a sentiment that is being shared by a growing number of researchers (Chuderski, 2014; Halford et al., 1998; Oberauer et al., 2008). The RMT appears to be an ideal exemplar of the relational integration hypothesis, given its remarkably simple concept and administration, and equally remarkable correlations with Gf.

Given the impressive RMT-related findings, it is perhaps surprising that it was not until Chuderski (2014) that a more formal analysis of the task was conducted to better understand the basis of these correlations. Chuderski manipulated the complexity of the relations to-be-considered by including a five-match condition (each relation involved binding five elements in the array for comparison, rather than the typical three) and by introducing a *different* match rule. The standard match rule, up to that point, required searching for identical stimuli (e.g., all digit strings end in the *same* last digit) whereas the *different* rule involved searching for distinct digits (e.g., all digit strings end in *different* last digits). The five-match condition produced an interaction effect with the *different* condition, such that performance dropped substantially more moving from three- to five-match with the different rule than with the same rule. Chuderski hypothesized that this was because non-identical digits could not be chunked the same way identical digits could be, leading to a much higher concurrent relational processing load moving from three to five digits to-be-integrated. The results of this manipulation strongly suggested that the task was primarily demanding relational integration. This was further supported by Chuderski finding no impact of visual interference (high interference involved arrays with many identical digits) on task difficulty. Together with the facts that (a) there is typically sufficient time to fully scan the array (over 5 seconds) and (b) not all stimuli are replaced when the array is updated (we henceforth refer to this feature as *string-preservation* meaning that some strings are preserved from trial-to-trial), this indicates that the task does not load heavily on attentional control. Chuderski again found an overall good correlation between the RMT and Gf (r = ~.41), though none of the experimental manipulations appeared to impact the magnitude of this correlation. Thus, it seemed that even the most basic form of the task could produce a valid measure of relational integration and, by extension, fluid intelligence.

The current study aimed to replicate and extend on our understanding of the features of the RMT that contribute to its success in predicting Gf. The end-goal is a clearer appreciation of the importance of relational integration in higher-order cognition. Three theoretically aligned RMT manipulations are investigated – cognitive complexity, attentional control, and inhibition. We consider each in turn.

Relational Complexity: To corroborate Chuderski's (2014) finding on match complexity, we also incorporated the *same* vs. *different* manipulation, but added an additional novel *ascending* condition. Before explaining the *ascending* manipulation, it is worth reiterating how *same* and *different* manifest in the task, what they convey theoretically, and why *ascending* may help round out the complexity manipulations. In the *same* condition, the match rule is "all strings within a row or column end in the same digit" while in the *different* condition, the match rule is "all strings within a row or column end in different digits". The *same* condition has a lower theoretical relational complexity (Halford et al., 1998) than the *different* condition because the first two end-digits in a same match [*same*(4,4,4)] can be systematically chunked together [*same*(4,4)], and distinguishing between these first two end-digits is not paramount to verifying whether the third end-digit (4) is also part of the relation – we need only know that the first two digits are the same and that both of them is (4). Contrarily, a different match [*different*(5,8,7)] cannot be chunked because, although together they can form the relation [*different*(5,8)], their unique identities must be kept available in order to verify that the third digit (7) is different from both the (5), as in [*different*(5,7)], and the (8), as in [*different*(8,7)]. Thus, according to the chunking principle in relational complexity theory (Halford et al., 1998), different matches requires more complex ternary relational integration than the binary relational integration involved in a same match. If we replicate earlier findings (Chuderski, 2014) demonstrating that the *different* match is more difficult than the *same* match, it would provide supporting evidence that the demands of the task may well lie in relational complexity. However, it is also possible that the two identical digits in the *same* condition are easier to chunk simply because they are identical, meaning lower-level visual identification strategies can be used for chunking, whereas the *different* condition necessitates higher-order relational integration. To clarify this, we included the *ascending* match condition: “at least one row or column has strings all ending in consecutively ascending digits”. Theoretically, an *ascending* match should have the same complexity as a *same* match (binary) because the first two digits can be chunked together. Consider the *ascending* relation [*ascending*(2,3,4)]. According to the relational systematicity principle (Halford et al., 1998, p. 808), the relational information between the (2) and (3) can be systematically chunked as [*ascending*(2,3)] because we do not need to know the difference between (2) and (3) in order to integrate the following relation [*ascending*(3,4)]. Rather, we need only know that both separate binary relations are ascending. To reiterate the earlier complexity analysis, this is different from the relation [*different*(5,8,7)] because each element (5), (8), (7) must be kept available in order to verify that each digit is different from both the other two digits. Thus, while both *ascending* and *different* have three unique elements involved in the relation, the effective complexity is only higher than *same* for *different* and not *ascending*. If the task is primarily demanding relational integration, we should see no difference in performance between *same* and *ascending*, but *different* should follow the same substantial drop in performance as in prior studies. Alternatively, a more linear decline between the three conditions (same > ascending > different) may indicate that both sources of demands are applying (visual identification and relational complexity).

Attentional Control (Scanning): Our second core experimental manipulation was on scanning demands. In the past, the RMT has always involved string-preservation where some of the nine strings present in the current array carry over to the next array (Chuderski, 2014; Krumm et al., 2009; Oberauer et al., 2008), reducing the amount of new information presented on each new array. Theoretically, this helps to minimize the amount of attentional control demands and maximize the relational integration demands. Operationally, the number of new stimuli that must be attended to is reduced and the primary demand remaining is in rapidly binding the strings into the target (match) relation. Although this task feature is theoretically meaningful, there has yet to be a clear experimental manipulation to determine how much this feature (attentional control) actually contributes to performance and to the relation with Gf. Kane et al. (2001) propose that the ability to actively maintain goal-relevant information in the face of irrelevant information is what connects WM tasks to Gf. Further findings by Kane and Engle (2003) on the Stroop task suggest that poor goal maintenance was a major factor in low WM participants struggling on WM tasks. Being frequently bombarded with arrays of completely new information in the RMT would require the ability to rapidly determine which strings are goal-relevant and efficiently dismiss irrelevant strings, on top of the relational integration demands already present in the task. To test the effects of attentional control through goal maintenance, our experiment includes both a *string-preservation* condition (some strings persist between arrays) and a *string-replacement* condition (all strings are replaced between arrays). The *string-preserve* condition and the *string-replace* condition should only differ in their association with Gf insofar as the attentional control demands of the task are related to Gf. In other words, if the *string-replace* condition significantly increases the relationship to Gf compared to *string-preserve*, it would indicate that attentional control is a significant component in Gf. To go one step further, if the task relies on *string-replace* to correlate with Gf, it would indicate that the RMT’s ability to predict Gf is being driven entirely by attention control, rather than the core relational integration demands of the task.

Inhibition (Visual Interference): Our final manipulation was to follow Chuderski’s (2014) work on visual interference which manipulated the number of identical digits in the array. In each of Chuderski’s arrays, one of the target string-ending digits was duplicated to non-string-ending positions. The theoretical idea was that these identical digits would act as distractors by increasing the similarity of targets (end-position digits) and distractors (non-end-position digits), demanding not just relational integration but the ability to inhibit distracting interference in the visual search process. In particular, they should adversely affect the *same* condition more than the *ascending* or *different* conditions because the identical digits are crucial to the *same* match but not either of the other two matches (Chuderski, 2014). Although Chuderski found no impact of interference on mean scores, one potential issue with his implementation of interference was that the high number of distractors (12 out of a possible 24 when excluding the progenitor’s string) could actually cue the participant to the target end-position digits, and thus cancel out any detrimental impact of the visual similarity. We explored this potential limitation by including three levels of interference with a similar target-duplication system: no interference had 0 digits duplicated, low interference had 6 duplicates (a novel condition), and high interference had 12 duplicates. We predicted that low interference, but *not* high interference, would produce a deficit in performance because it causes some visual interference without overtly cueing the participant to the target digits. Inhibition is also related to attentional control (Engle, 1996). Although inhibition usually refers to associative activation in long-term memory (Hasher & Zacks, 1988), explicit visual inhibition caused by visually identical stimuli can also have a strategic component related to task performance, in purposeful avoidance of the allocation of attention to distracting elements (Lu et al., 2017). Thus, a difference between the interference conditions in predicting variance in Gf can still represent the contribution of visual inhibition in the RMT’s relationship to Gf.

5.2.1. Aims and hypotheses

The current study was conducted to systematically manipulate task features of the *Relation Monitoring Task* to determine what makes the task so good at predicting Gf over-and-above classic “store-and-process” WM measures. We extend on Chuderski’s (2014) manipulations by further investigating the roles of attentional control demands through an elaborated visual interference manipulation and by comparing string-preservation with string-replacement. We also further manipulated complexity with the addition of the *ascending* match type to remove the confound of identical digits contributing to lower-order visually-oriented chunking. To provide a stronger conclusion in determining what the RMT shares with Gf, we included three Gf measures, so we could form a latent Gf measure. We also included two classic criterion measures of WM: a complex-span and an *n*-back. It was predicted that the extent of the relationship with Gf will be largely determined by the capacity of the RMT to measure relational integration (the *relational integration hypothesis*). Specifically, it was hypothesized (H1) that the *different* condition will increase the difficulty of the task and the relationship to Gf compared to *same* and *ascending*, as *different* matches require a higher relational complexity to integrate; while *same* and *ascending* have the same theoretical complexity, and so should have similar performance. It was also hypothesized (H2) that *string-replace* will add an additional unique component in predicting Gf over *string-preserve*, in line with the additional attentional control demands required in dealing with a full array of new strings. However, in line with the relational integration hypothesis, both versions of the task will predict Gf (i.e., string-replace is not necessary for the RMT to predict Gf). Finally, it was hypothesized (H3) that *low interference* but not *high interference* will decrease performance on the task, because it visually interferes with participants without overtly cueing them to the target. Again, in line with the relational integration hypothesis, all versions of the task will predict Gf. In summary, in all cases, we predict that the relational integration demands of the RMT will predict Gf over-and-above the two criterion WM measures, which primarily measure classic "store-and-process" demands; and further increases will be concomitant with the respective theoretical demands of the manipulations.

5.2. Method

5.2.1. Participants and Procedure

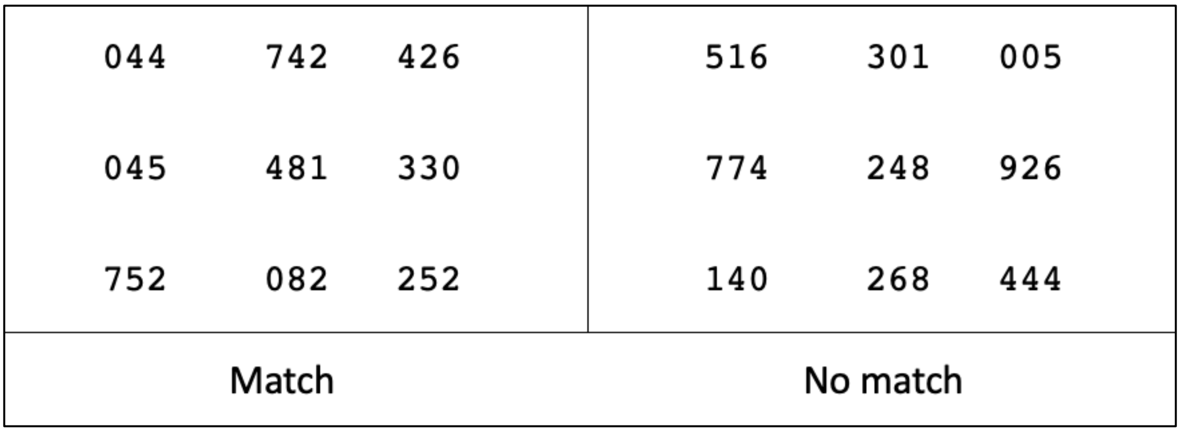
A total of 105 participants took part in exchange for course credit. Five participants were excluded due to unacceptably low scores on at least one measure, indicating they did not understand the task instructions or were purposely not engaging in the task.[11] Of the remaining 100, 67 were female and 33 were male, with an average age of 19.47 (SD = 2.12). Participants undertook six tasks: the RMT, three measures of Gf (APM, Letter Series, Latin Square Task), a complex-span (operation-span), and an *n*-back (spatial *n*-back). Participants completed the tasks in a random order in 90-minute sessions, in groups of up to eight in computer labs at the University of Sydney.

5.2.2. Measures

Relation Monitoring Task (RMT)

The RMT involved presenting a continuous 3 x 3 array of 3-digit number strings. The task was to respond (with the spacebar) whenever an array matching the current match rule was presented (see Figure 5.1). If the array did not match the current rule, the participant was to wait for the next array, which would replace some or all strings (depending on the condition) with new ones. Each array was presented for 5.5 seconds with a 100ms interval.

[11] Of the five participants excluded, two scored 0 for the LST-DC, two scored 0 for the Letter Series, and one scored 1 for APM (all these scores were more than 3 SDs below their respective means, with distribution plots demonstrating clear outliers). All three tasks included items ranging in difficulty, including particularly easy items that are expected to be trivial for university-level adults.

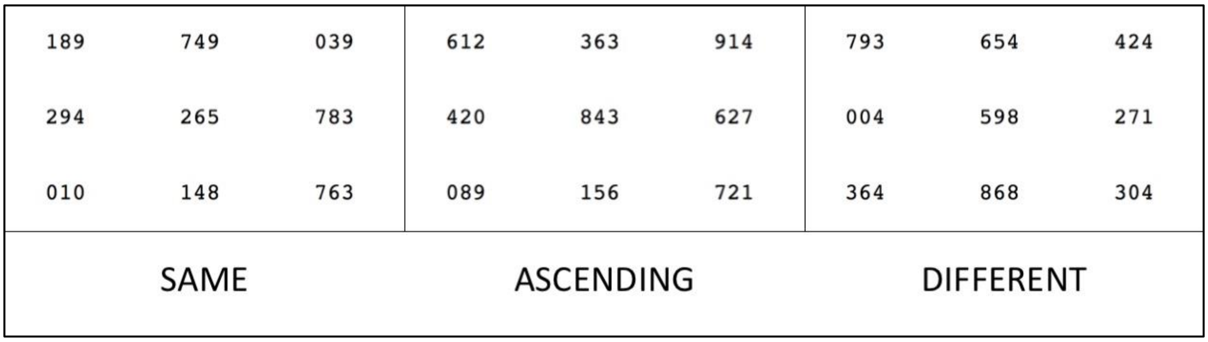


044	742	426	516	301	005
045	481	330	774	248	926
752	082	252	140	268	444
Match			No match		

Figure 5.1. Examples of two arrays with the ‘same’ match rule from the Relation Monitoring Task. In the ‘match’ example (left), all three number strings in the bottom row end with the same digit, 2. In the ‘no match’ (right), there are no rows or columns where all three number strings end in the same digit.

There were three experimental manipulations, balanced across one-another: complexity (same/ascending/different), string-preservation (string-preserve/string-replace), and interference (none/low/high). Each manipulation is detailed in the paragraphs below. Participants completed a total of six blocks, with a unique complexity and string-preservation combo: *same-replace, same-preserve, ascending-replace, ascending-preserve, different-replace,* and *different-preserve*. Each block had 36 test trials, half of which were matches. Score was derived through the proportion of correct hits on match trials minus the proportion of false alarms on no-match trials (e.g., 15/18 correct matches and 4/18 incorrect false alarms would lead to a score of (.83 - .22 =) .61 for that block). The three levels of the interference condition (none/low/high) were balanced within each block.

RMT: Complexity. The three match rules (representing complexity) are demonstrated in Figure 5.2. The *same* condition involved matches where three strings in a row or column ended in the same digit. The *different* condition involved matches where three strings in a row or column ended in different digits. For the new *ascending* condition, the match was whenever three strings in a row or column ended in consecutively ascending digits. Participants were given instructions and practice on each match type, including specific instructions on the *ascending* condition that made it clear that the ascending digits must be in consecutive order (top-to-bottom for columns, left-to-right for rows). A reminder of the current match rule was always present to the left of the array.

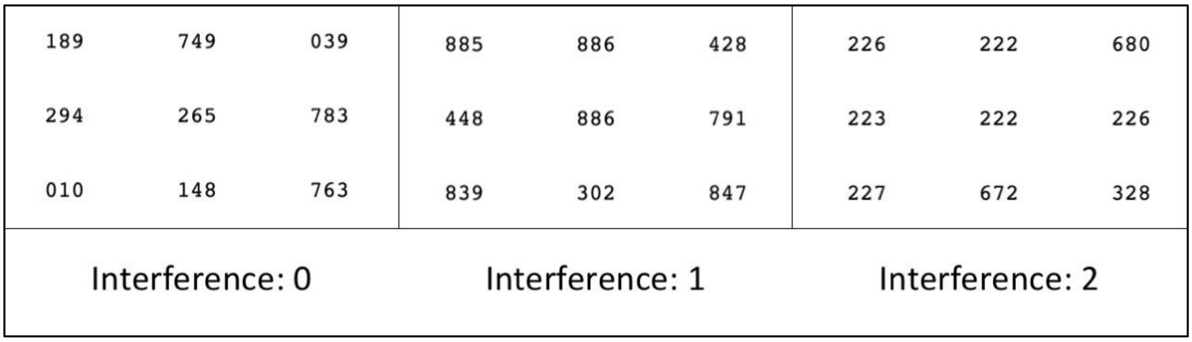


189	749	039	612	363	914	793	654	424
294	265	783	420	843	627	004	598	271
010	148	763	089	156	721	364	868	304
	SAME			ASCENDING			DIFFERENT	

Figure 5.2. Examples of match arrays for each complexity condition. Same: at least one row or column with strings all ending in the same digit (match: top row). Ascending: at least one row or column with strings all ending in consecutively ascending digits (match: top row). Different: at least one row or column with strings all ending in different digits (match: middle row).

RMT: String-preservation. The string-preservation parameter was manipulated by comparing the score of *preserve* blocks against *replace* blocks, averaged across complexity. In *preserve* trials, 1-4 strings (at random) persisted from one array to the next – this replicated Chuderski’s (2014) methodology. In *replace* trials, all strings were always replaced with new ones on each new array.

RMT: Interference. The final manipulation was interference with three levels: *Int-0, Int-1,* and *Int-2* (corresponding to none, low, and high, respectively). These levels are demonstrated in Figure 5.3. *Int-0* were regular trials with no duplicated digits. *Int-1* caused one random string-ending digit to duplicate six times across the array into non-string-ending positions. *Int-2* was similar, except the progenitor digit duplicated twelve times. In match trials, the progenitor digit was always part of the target match, while in non-match trials, it was a random string-ending digit. *Int-0* and *Int-2* replicated Chuderski (2014), while *Int-1* was a novel addition. Each level of interference was presented an equal number of times per block such that for each of the 18 matches and each of the 18 non-matches in a block, there were six *Int-0*, six *Int-1*, and six *Int-2* arrays, distributed randomly amongst complexity and string-preservation.



189	749	039	885	886	428	226	222	680
294	265	783	448	886	791	223	222	226
010	148	763	839	302	847	227	672	328
Interference: 0			Interference: 1			Interference: 2		

Figure 5.3. Examples of arrays with interference manipulation. 0: no digits replaced. 1: six digits replaced by a random string-ending digit. 2: twelve digits replaced by a random string-ending digit. Int-1 shows a low level of visual overlap caused by the duplicated 8, while Int-2 shows a high level of visual overlap caused by the duplicated 2.

Raven's Advanced Progressive Matrices

Participants completed the abbreviated 20-item version of Raven's APM (odd items + items 34 and 36) as an indication of Gf, as in earlier chapters. Participants had 20 minutes to solve as many items as possible.

Letter Series

Participants had four minutes to complete as many of 15 Letter Series items as they could. Each item involved a patterned sequence of letters followed by an underscore to indicate that the task was to complete the pattern by inserting a single letter to the end of the sequence. Like the APM, the items become progressively more difficult.

Latin Square Task

The Latin Square Task (LST) was employed as an additional criterion measure. In the LST, participants are presented with an incomplete 4 x 4 matrix, partially filled with four types of shapes (a circle, square, triangle, and cross) and including one target '?' cell. Participants are informed of the one defining rule of the LST: that each row and column may only contain *one* of each of the four shapes from the set of shapes. The task is to determine which of the four types of shapes should be in the marked target cell. Items primarily vary in difficulty through complexity (how many rows and columns must be considered to derive the target cell). Based on task analyses in Chapters II and III, the LST is thought to relate more to relational WM than to Gf, because the only rule in the task is given.

For this implementation of the task, we administered 24 items split evenly by complexity. Half of these items were standard LST items while half were *dynamic completion* (DC) items (see Chapter III), where participants could dynamically fill non-target cells of the matrix as they solved for the target cell. This manipulation was not central to the current study, so the two types of items are collapsed over.

Operation Span

Participants completed the OSPAN with set sizes of 3, 4, 5, and 6 (two sets of each). In each set, participants alternated between memorizing a letter and verifying the truth of a mathematical operation. Once all letters for that set had been presented, participants attempted to recall the letters in the order they were presented. Scores were calculated as total number of correct letters recalled (*OSPAN Letters*) rather than the number of correct letters in fully recalled sets (*OSPAN Capacity*). The partial scoring of OSPAN Letters is preferred because it accounts for the same variance picked up by absolute scoring with OSPAN Capacity, but also accounts for additional variance that would be otherwise discarded (Redick et al., 2012).

Spatial n-back

Participants completed a spatial version of the *n*-back with two blocks of 2-back and two blocks of 3-back. In each block, participants were presented with a 3 x 3 cell matrix. Every two seconds, a blue square would flash for one second inside a random cell. The participant’s task was to respond whenever a blue square appeared that was on the same cell as a blue square from *n* steps back (i.e., 2 squares back on the 2-back condition). Score was derived as number of hits minus number of false alarms, then averaged across the four blocks.

5.3. Results

5.3.1. RMT Manipulations: Performance effects

Descriptives for the student sample are presented in Tables 5.1 and 5.2. The six RMT blocks demonstrated acceptable internal consistency, $\alpha = .79$, despite differences in the match complexity of the conditions, which a repeated-measures ANOVA determined to be significant, $F_{2,198} = 222.11$, $p < .001$. Two planned contrasts revealed the matches with lower relational complexity (*same* and *ascending*) had higher performance than the match with higher complexity (*different*), $F_{1,99} = 236.90$, $p < .001$, $\eta_p^2 = .71$; yet the same condition also had higher performance than the ascending condition, $F_{1,99} = 201.93$, $p < .001$, $\eta_p^2 = .67$.

Table 5.1. *Descriptives for all measures, with RMT split by match complexity and preservation.*

Measure	Mean	SD
Raven's APM	0.61	0.19
Letter Series	0.69	0.12
OSPAN (Letters)	0.83	0.18
N-back DV*	2.42	1.64
Latin Square Task	0.82	0.16
LST-Basic	0.80	0.18
LST-DC	0.83	0.18
RMT Grand Total	0.67	0.13
RMT Same Total	0.87	0.11
RMT Same	0.86	0.14
RMT Same-P	0.87	0.11
RMT Asc Total	0.64	0.18
RMT Asc	0.62	0.20
RMT Asc-P	0.65	0.21
RMT Diff Total	0.50	0.20
RMT Diff	0.50	0.22
RMT Diff-P	0.51	0.24

**Note:* N-back DV reflected as raw score (block average hits minus false alarms) as task is not itemized. RMT conditions marked with '-P' indicate the preserve variant, while those without the suffix indicate the replace variant.*

For interference (digits duplicated), there was no main effect on performance in a repeated-measures ANOVA, $F_{2,182}$ = 1.24, p > .05, indicating that neither low nor high interference decreased performance.

Table 5.2. *Descriptives for RMT split by interference.*

RMT Condition	Mean	SD
RMT Grand Total	0.67	0.13
Interference 0	0.65	0.16
Same Int 0	0.84	0.16
Asc Int 0	0.61	0.22
Diff Int 0	0.50	0.25
Interference 1	0.67	0.14
Same Int 1	0.87	0.12
Asc Int 1	0.63	0.20
Diff Int 1	0.51	0.23
Interference 2	0.66	0.14
Same Int 2	0.87	0.11
Asc Int 2	0.63	0.22
Diff Int 2	0.46	0.24

5.3.2. RMT Prediction of Gf: Controlling for Working Memory

The purpose of this set of analyses were to verify that the RMT correlated with Gf over-and-above the two criterion WM measures, complex span and *n*-back. An initial *Gf* factor was derived through principal axis factoring with varimax rotation on the two Gf measures, APM (α = .80), Letter Series (α = .75); and the LST (α = .76[12]). As discussed in Chapters II and III, I acknowledge that the LST may lie somewhere short of Gf, so the analyses are repeated further on without the LST also. With the LST, this factor accounted for 65% of the variance in the three measures with an eigenvalue of 1.95; with factor loadings of .774 for LST, .691 for APM, and .602 for Letter Series.

As demonstrated in Figure 5.4, the RMT had a considerable *r* = .61 with *Gf* – this is in comparison with *r* = .41 reported by Chuderski (2014). As seen in Table 6.3, the *n*-back also correlated with *Gf* (*r* = .53) but the OSPAN did not. As seen in past research (Redick & Lindsey, 2013), the *n*-back and OSPAN also did not correlate with each other (*r* = -.02).

[12] The LST reliability here is derived through three complexity subscales (2D/3D/4D) averaged across basic and DC LST variants. This produces a lower-bound estimate of the total scale α but is comparable to the LST total *α* = .79 reported by Birney et al. (2012) in a population of managers.

Table 6.4 provides the full correlation matrix separating RMT conditions and tasks. Contrary to Redick et al.'s (2012) recommendation, using OSPAN Capacity rather than OSPAN Letters generally increased the OSPAN's correlations across the board, though it remained the weakest predictor of *Gf* (*r* = .25). As per Redick et al.'s suggestion, the following regression analyses will continue to use OSPAN Letters. Regardless of whether OSPAN Letters or OSPAN Capacity is used as a predictor, the outcomes of the analyses do not change.

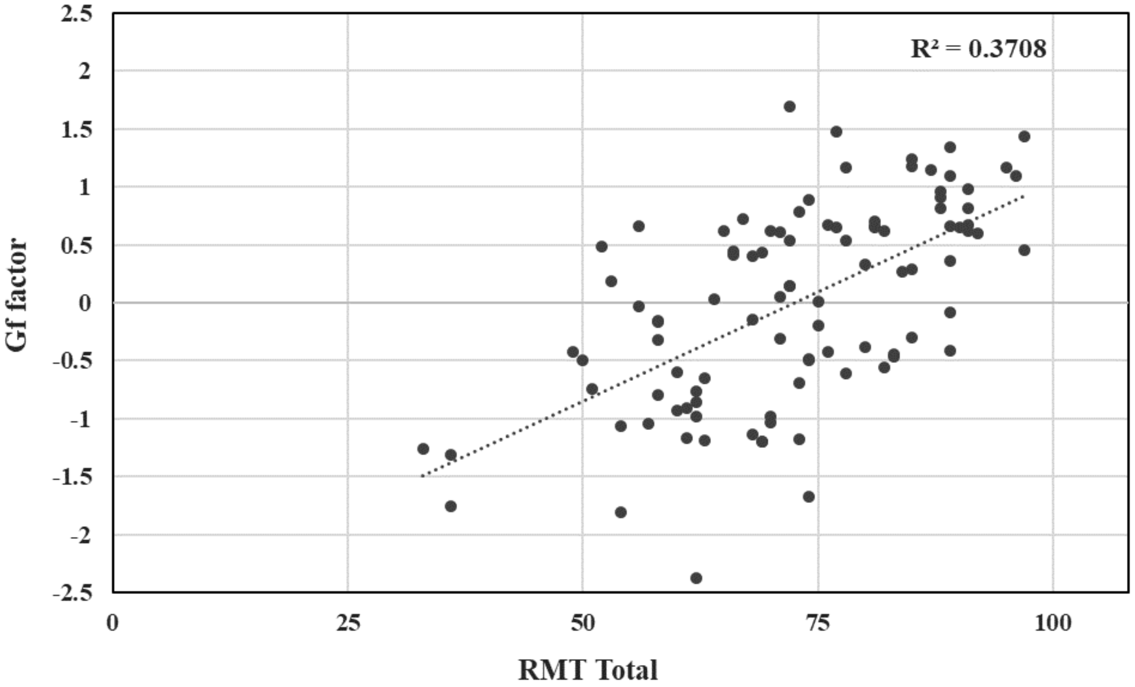


Figure 5.4. Scatter plot with RMT total score (raw; out of 108) on X-axis and *Gf* factor on Y-axis.

Table 5.3. *Correlation between WM measures and Gf factor*

	RMT	OSPAN	N-back
Gf Factor	.61**	.17	.53**
RMT Total	-	.13	.44**
OSPAN (Letters)		-	-.02

Table 5.4. *Full condition and task correlation matrix.*

	RMT (T)	RMT-Pres	RMT-Repl	RMT-Same	RMT-Asc	RMT-Diff	LST	APM	L-Series	N-back	OSPAN
RMT (Total)	-										
RMT-Preserve	.93**	-									
RMT-Replace	.93**	.73**	-								
RMT-Same	.75**	.65**	.74**	-							
RMT-Ascending	.81**	.78**	.74**	.47**	-						
RMT-Different	.86**	.80**	.79**	.54**	.47**	-					
LST	.51**	.51**	.43**	.39**	.43**	.42**	-				
APM	.47**	.51**	.36**	.35**	.37**	.41**	.54**	-			
L-Series	.53**	.51**	.47**	.50**	.45**	.38**	.47**	.42**	-		
N-back	.44**	.44**	.37**	.50**	.26**	.38**	.47**	.32**	.51**	-	
OSPAN (Letters)	.13	.18	.06	.04	.09	.15	.06	.24*	.15	-.02	-

** Significant at < .05; ** Significant at < .01.*

RMT = Relation Monitoring Task. LST = Latin Square Task. APM = Raven's Advanced Progressive Matrices. L-Series = Letter Series. OSPAN = Operation Span (Letters).

More important to our research question was to determine if the RMT's relationship to *Gf* was demanding similar processes as the classic WM measures (*n*-back and OSPAN), or if it was indeed contributing its predicted variance over-and-above these typical WM measures. We conducted a multiple linear regression predicting the *Gf* factor with the first model containing the two classic WM measures and the second adding the RMT. As seen in Table 5.5, the classic WM measures predicted a considerable 31% of variance in *Gf*, mainly driven through the *n*-back (sr^2 = .28, p < .001). The OSPAN also predicted a significant, though small, unique portion (sr^2 = .03, p < .05). Importantly, once we added the RMT, the predicted variance increased to 48%, a significant change, ΔR^2 = .166, p < .001. With the RMT in the model, the OSPAN now provided nothing unique, with its predicted variance of *Gf* subsumed by either the *n*-back or RMT. The *n*-back maintained some unique predictive variance of *Gf* (sr^2 = .09, p < .001) though the RMT had the highest unique component predicting *Gf*, sr^2 = .17, p < .001.

Table 5.5. *Multiple linear regression with the two classic WM measures predicting Gf (Model 1) then adding RMT (Model 2)*

Model	Measure	β	sr^2	R^2	ΔR^2
1: Classic WM Measures	OSPAN (Letters)	.185	.03*	.310	.310**
	N-back	.529	.28**		
2: Add RMT	OSPAN (Letters)	.122	.01	.476	.166**
	N-Back	.327	.09**		
	RMT	.459	.17**		

Although the LST has been used as a *Gf* measure (Birney et al., 2012), it was primarily designed to tap relational integration (Birney et al., 2006). As discussed in Chapters II and III, the lack of rule induction demands means the LST is unlikely to qualify as a full Gf task. Thus, it is possible that the strong relationship[13] between the RMT and our *Gf* factor is primarily a result of the LST being included in the *Gf* factor. To demonstrate that the relationship still holds without the LST, we reconducted the prior regression, this time predicting the common factor formed only from APM and Letter Series, using the same extraction method. This two-task *Gf* factor accounted for 70.9% of variance in the two component measures with an eigenvalue of 1.42. The results were largely unchanged from the three-task *Gf* regression (Model 1 $R^2 = .334$; Model 2 $R^2 = .460$). The only substantial change was that the OSPAN remained a significant unique predictor in the second model ($sr^2 = .03$, $p = .02$), though it was still the lowest of the three tasks (RMT $sr^2 = .15$, $p < .001$; *n*-back $sr^2 = .08$, $p = .001$). Thus, the strong relationship between the RMT and *Gf* observed here does not appear to be inflated simply due to the inclusion of the LST in the *Gf* factor. Given the largely identical outcomes between the two-task and three-task *Gf* factors, we proceed with the remaining analyses using the three-task *Gf* factor. However, this does mean it may be more appropriate to think of the *Gf* factor as more of a relational integration factor than a full *Gf* factor *per se*, simply because one of the tasks involved does not capture the rule induction demands unique to Gf.

5.3.3. RMT Prediction of Gf: Experimental manipulations

The first regression analysis made it clear that the RMT does indeed have an impressive relationship to *Gf*, accounting for 16.6% of Gf variability over-and-above classic WM measures. Although this is substantially higher than Chuderski's finding of 5.9%, it should be noted that he included additional WM measures. Our next regressions (which are the novel component of our experiment) aimed to uncover the parameters involved in the

[13] To the best of our knowledge, this is the highest correlation between the RMT and Gf observed in published research yet.

RMT that are substantive to this relationship. These include complexity (match type), inhibition (interference), and attentional control (string-preservation).

For match complexity, we regressed (in order) *same*, then *ascending*, then *different*. The first model, containing just *same*, accounted for 24% of the variance in *Gf*, $R^2 = .244$, $p < .001$. Adding *ascending* increased this to 33%, $\Delta R^2 = .086$, $p = .001$. Adding *different* then further increased this to 38%, $\Delta R^2 = .044$, $p = .012$. In this final model, all three predictors held small but significant unique contributions (same: $sr^2 = .04$, $p < .05$; ascending: $sr^2 = .04$, $p < .05$; different: $sr^2 = .04$, $p < .05$) while still leaving the majority $R^2 = .26$ as shared variance.

For interference, we conducted similar analyses, iterating on the regression model as the task increased in interference. It is worth reiterating that there were no mean differences found between the interference conditions. The following results are thus particularly interesting. The first model, with only no-interference trials, accounted for 25% of the variance in *Gf*, $R^2 = .249$, $p < .001$. The second model added low interference trials, and increased the explained variance in *Gf* to 49%, $\Delta R^2 = .240$, $p < .001$. However, the third model adding high interference trials, did not increase the variance explained in *Gf* significant, $\Delta R^2 = .006$, $p > .05$. In the final model, only the low interference provided a unique contribution, $sr^2 = .21$, $p < .001$.

Our final regression model considered the string-preservation parameter. Again, it is worth keeping in mind that string-preservation also had no impact on mean scores. The first model consisted solely of string-preserve trials (which theoretically minimizes attentional control demands) and accounted for a significant 26% of the variance in *Gf*, $R^2 = .259$, $p < .001$. Adding the string-replace trials (which theoretically translates to higher attentional demands) increased this accounted variance to 39%, $\Delta R^2 = .13$, $p < .001$. In the final model, only string-replace trials had a significant unique contribution, $sr^2 = .13$, $p < .001$.

5.4. Discussion

The aim of the current study was to experimentally manipulate the RMT as a measure of relational integration by demanding different levels of relational complexity, attentional control, and inhibition to determine what task features are essential for the task to produce its impressive prediction of Gf. Overall, our results were consistent with prior research demonstrating a significant correlation between RMT and Gf (Chuderski, 2014; Krumm et al., 2009; Oberauer et al., 2008) and in fact, our RMT showed an even stronger correlation (*r* = .61) than prior findings (in the *r* = .3~.5 range). It is worth reiterating how remarkable such a powerful relationship is in individual differences research (J. Cohen, 1988; Gignac & Szodorai, 2016), particularly when considering the apparent simplicity of the RMT, which requires no explicit storage over time or advanced mental manipulation. This simple task can predict as much as 37% of variance in a latent *Gf* factor composed of advanced, abstract series completion tasks such as Raven’s, Letter Series, and the LST. Theorizing surrounding the RMT seems to indicate this result simply comes about due to the purity of the task in measuring a most fundamental aspect of WM: relational integration (Halford et al., 1998; Oberauer, 2009a). Our novel experimental manipulations illustrated that – in line with the relational integration hypothesis – all versions of the task could predict Gf. Although the majority of predicted variance in Gf was shared amongst the different RMT conditions, we did identify further unique components of the RMT related to increases in relational complexity and additional attentional control and inhibition demands.

Our three RMT match conditions (*same, ascending, different*) appeared to be tapping a similar demand in WM – which is theorized to be relational integration – a conclusion emerging from high reliability and a large amount of shared variance between the complexity conditions accounted for in Gf (the match regression findings suggest over two-thirds of the variance was shared between conditions). Beyond this shared variance, and contrary to expectations, all three levels of complexity provided something unique in predicting Gf. Although this was consistent with the mean differences (in that each match was more difficult than the last), our hypothesis was that *ascending* would offer nothing unique in predicting Gf over-and-above *same* because it has the same theoretical complexity (binary), because the first two elements in the series can be systematically chunked, unlike in a *different* match. Although *same* and *different* did indeed each have independent components related to Gf, *ascending* also did, indicating there may be some unique demand related to the ability to apply systematicity (Halford et al., 1998) to the relational integration process for elements which are different in appearance, but can be systematically chunked down. For instance, once the relation between the digits (4) and (5) has been verified as ascending, they can be systematically chunked down into a single binding [4,5] and ascending relation needs only to know that the next digit follows the order as (6). We hypothesized *ascending* required no additional demand over *same* because they both require sequential instances of binary relational integration (as opposed to *different*, which requires ternary relational integration). Although this may still be the case, our results indicate that the added challenge of applying systematic chunking to two visually distinct digits (4,5) may constitute a demand related to both performance and Gf. It is also possible that this unique *ascending* demand came about through the restriction on scanning: because the ascending matches were always consecutive and sequential, the matches were most easily checked by scanning left-to-right and top-to-bottom. Although they could be scanned in opposite directions, it would require reversing the match being checked against to *descending*. Conversely, for *same* and *different* matches, participants could scan right-to-left or bottom-to-top with only the one rule.

For the interference manipulation, again, the majority of variance explained in Gf was shared amongst the three conditions (*Int-0, Int-1, Int-2*). The interference levels were virtually indistinguishable on a mean difference level however, the low interference (*Int-1*) provided a considerable, unique component in predicting Gf which was hypothesized to be the demand of dealing with additional attentional interference of multiple duplicated digits. Our actual hypothesis related to mean differences, in that high interference (*Int-2*) may have been cueing participants to the target match, while low interference represented a ‘sweet spot’ of interfering, but not cueing. This sweet spot still does not make an apparent difference in task difficulty (Chuderski, 2014), but it does appear to tap a unique demand related to Gf, independent from relational integration. Although such a demand would exist independently of the relational integration hypothesis, it could be explained by a visual search strategy where participants purposely allocated no attention to potential distractors (Lu et al., 2017) – the non-string-ending digits. The finding that this strategy could too relate uniquely to Gf is preliminary but plausible, given that tasks such as Raven’s often involve many distinct visual elements which must be considered independently across rows and columns (Verguts & De Boeck, 2002) and then ruled out as irrelevant (inhibited) or maintained for further consideration as appropriate (Carpenter et al., 1990).

Our final manipulation, string-preservation, is perhaps the most important. It is a parameter often taken for granted yet with potentially critical implications concerning the role of attentional control in the RMT. Prior work with the RMT has included string-preservation (Chuderski, 2014; Krumm et al., 2009) to minimize the amount of new scanning required, thus maximizing the relational integration demands while minimizing the attentional control demands. Our results indicate that string-preservation (like interference) has no impact on overall task performance but does significantly change the relationship with Gf. That is, *string-replace* trials offered a unique contribution that substantially increased the relationship to Gf (accounting for exactly one-third of the variance in the RMT when compared to *string-preserve* trials). This means that, in line with the relational integration hypothesis, the task functions perfectly well as a pure predictor of Gf with string-preservation, but the relationship to Gf can be enhanced further by adding the incremental attentional control demands reflected in string-replacement, where rapid, flexible binding and unbinding is relevant.

It is also worth reiterating that the RMT surpassed classic WM measures, predicting substantial variance over-and-above the unique and shared variation accounted for by complex-span and *n*-back. It has been frequently theorized (Krumm et al., 2009; Oberauer et al., 2008) that this is because the RMT taps a fundamental aspect of WM: relational integration, which is also captured (albeit impurely) in these traditional WM measures (Oberauer et al., 2008), which may instead more strongly reflect passive storage or updating components of WM (Chapter IV). Like Chuderski (2014), we contributed further evidence to the relational integration hypothesis – the suggestion that Gf can be most fundamentally captured by measuring relational integration – by finding that all experimental variations of the RMT tap a similar demand consistent with the ability to rapidly establish bindings between independent elements.

Some future suggestions should be considered. On the topic of string-preservation, there is scope to further assess its impact on task demands. In our study, we replicated Chuderksi’s methodology of preserving 1-4 strings at random and contrasting it to our novel manipulation of replacing all the strings (i.e., preserving none of them). However, in Oberauer et al. (2008) and Krumm et al. (2009), only a single string was preserved between trials, but the task updated at a faster rate (every 2 seconds). In contrast to this 8-string preservation, our manipulation seems minor, and yet still made a significant unique contribution in the relationship with Gf after controlling for WM, with the 1-4 string-preservation accounting for about one-third of the effect. That such a seemingly minor manipulation had such an impact indicates that comparing a wider range of string-preservation (i.e., up to 8 strings, rather than 4) could elucidate a further substantive demarcation of relational integration and attentional control demands. It is possible that although the increase in strings preserved (up to 8) may further minimize attentional control demands, but the 2-second response window may counteract this. A future task analysis could thus consider both string-preservation and response window independently.

5.4.1. Conclusion

In this experiment, we found encouraging results for the *Relation Monitoring Task* as an assessment of relational integration and predictor of fluid intelligence. Theoretically, the RMT is a task demanding relational integration but, functionally, it appears to be a powerful, reliable predictor of Gf. This is perhaps the most important implication of our results. Our battery of abbreviated Gf tasks took approximately 60-75 minutes to administer. A full battery typical of recruitment assessment can take 4-8 hours (Chuderski, 2014) or even several days (Robertson, Gratton, & Sharpley, 1987). Yet the RMT, which takes only about 20 minutes to administer, predicts as much as 37% of variance in Gf – a correlation so high it is only seen in 2-3% of individual differences studies (Gignac & Szodorai, 2016).

In summary, we replicated prior research demonstrating a powerful relationship between the RMT and more theoretically complex Gf measures. The RMT is an insightful task because it requires no explicit storage over time and no advanced mental manipulation, instead primarily measuring relational integration. We continued Chuderski's (2014) breakdown of the task, supporting the notion that the task is a measure of the ability to rapidly establish bindings between multiple elements for relational integration. For the first time, we have also demonstrated that the task appears to have some attentional control demands associated with it, though these are not crucial to its relationship with Gf. Our results are thus strong evidence for the relational integration hypothesis (Bateman, Birney, & Loh, 2017; Chuderski, 2014; Halford et al., 1998; Oberauer et al., 2008) but may also coincide with a more attentionally-oriented perspective (Kane et al., 2001; Shipstead et al., 2016). Our findings support both theories but suggest that each may serve a different purpose in the prediction of Gf. Maintaining focus during a complex task (such as Raven’s) and orienting attention towards the goals of the item are helpful but represent a fundamentally different demand to the crucial ability to integrate a relation by binding elements in a mental workspace such as working memory. Ultimately, no matter how focused and well-oriented one is to the goals of the task, the capacity for relational integration can prove to be a cognitive obstacle only overcome with the capacity to strategically and systematically chunk (Halford et al., 1998). Abstract reasoning and Gf are certainly complex constructs, with prototypical tasks that tap a wide range of theoretically elusive cognitive demands. Latent variable analysis is the current gold standard for unravelling this constellation of demands but theoretically-driven experimental manipulations are key to determining what cognitive demands are most essential for inter-individual variation. The importance of Gf tasks in applied settings such as recruitment and aptitude highlight a need both to understand these cognitive demands and to consider how we can assess them in a way that is both cost- and time-effective. The RMT appears to be a task that can answer the theoretical questions on the source of demands and provides a pragmatic substitute for large-scale Gf batteries.

VI. STUDY 4: THE SWAPS TASK

Working memory is a critical system that allows us to maintain information in a highly accessible state for further processing. Theories of working memory must answer three critical questions: what limits working memory, how these limitations vary between individuals, and how they can explain the link between working memory and fluid intelligence (Conway et al., 2007). Of these theories, two prominent perspectives that we have focused on throughout this thesis are theories of binding capacity (Oberauer, 2009) and theories of attentional control (Engle & Kane, 2004). Although both provide answers to the three critical questions, it has been difficult to demarcate the two perspectives experimentally, as they often share predictions and explanations. Thus far, we have found indications that binding capacity in relational integration appears to provide a more parsimonious answer to the third question – linking WM to Gf. In each of the three prior chapters, the relational integration explanation has accounted for the WM-Gf overlap while attentional control seems to be a supplementary but non-essential component. However, because attentional control is so broad, it is still difficult to fully distinguish from binding capacity. For instance, although increased binding demands in *access-random* in the ACT (Chapter IV) appear to support relational integration theories, the random ordering may also require more attentional control to keep more bindings active. In Chapter V, attentional control demands appeared relevant but not critical. However, the task in itself (the RMT) was designed to measure relational integration, so it possible that the binding interpretation was getting some benefit from the base task overlapping with Gf through positive manifold. For instance, if participants are motivated to do well on all tasks, then some correlation will appear between RMT and Gf just through that motivation and irrespective of any actual correlation the tasks have. In our analyses on the RMT, this motivation would have been subsumed into the binding interpretation. The study for this chapter was developed to provide a direct comparison between binding capacity and attentional control demands by independently manipulating them in a single task, the Swaps task (Crawford, 1988). Both binding capacity and attentional control demands are expected to contribute substantially to performance in the Swaps task (something that was not seen in the RMT, where attentional control demands modestly influenced performance). The Swaps task involves mentally rearranging a series of letters according to a set of simple instructions. Thus, it also exemplifies the key difference found in the ACT between *access-fixed* and *access-random*. The rearrangement exclusive to *access-random* was thought to load highly on binding capacity, because participants must juggle three independent bindings rather than one systematic binding. The Swaps task removes the arithmetic and access/retention components and focuses completely on this rearrangement. The remainder of this introduction explains the Swaps task, then describes how (i) our novel manipulation of Letters and (ii) the established manipulation of Steps represents manipulations in binding capacity and attentional control, respectively.

6.1. Introduction to the Swaps Task

In the *Swaps Task* (Crawford, 1988; Stankov, 2000; Stankov & Crawford, 1993), participants follow instructions on a screen that direct them to 'mentally swap' the order of three letters. For instance, the letters [J K L] may be presented simultaneously, accompanied by lines of instructions such as *"Step 1: Swap 1 and 3 | Step 2: Swap 2 and 3"*. To this item, participants should respond with [L J K] (because Swap 1: J K L to L K J; Swap 2: L K J to L J K) to answer correctly. This simple task is useful because the premise of the task is straightforward (requiring few instructions) and yet cognitive demands can be easily manipulated under the same rule scheme. There are two primary demands identified in the task. There are active storage demands in having to hold interim solutions over the course of the problem (across steps). There are also binding demands in having to unbind and bind letters to new positions in the order. In the original studies on the Swaps (Stankov, 2000;

Stankov & Crawford, 1993; Stankov & Schweizer, 2007), the primary manipulation of an increased number of swaps (i.e., Steps) both decreased performance and increased the relationship of the task to the more complex Gf measures. Because the actual number of elements involved in the problem remained fixed at three letters, the actual binding capacity demands remained consistent swap-to-swap (two being changed; and three in total). Instead, it is more reasonable to suggest that WM is being strained through the attentional control demands in having to maintain the current iteration of the order from swap to swap. Although this could also be related to a build-up of proactive interference making it progressively more difficult to maintain each binding from swap to swap (which would implicate binding capacity demands) (for a review, see Oberauer, Awh, & Sutterer, 2017), the performance trajectory of increasing steps is primarily linear (see the next section for details) rather than exponential. Thus, when Stankov (2000) finds that the linear effect of Steps covaries with Gf, it is reasonable that he also concludes that Gf is related to attentional control demands. However, there are reasons to suggest that this conclusion may be incomplete. The following sections present a task analysis on the Swaps task, first analysing the Steps manipulation, then introducing the novel Letters manipulation.

6.1.1. Steps

The classic manipulation of difficulty in the Swaps task is to increase the number of steps. For instance, rather than requiring only two swaps to reach the solution (two steps), a more difficult problem may require three or even four swaps. Stankov (2000) found that the increase in Steps produced a smooth decrease in accuracy as steps increased, with the average percentage correct falling by approximately 8% for each additional step (from 90% at 1-step to 66% at 4-step). Although there was not a large enough range of steps in Stankov’s (2000) study to determine if this decrease plateaus, Bowman (2006) then provided additional data with steps ranging up to eight. Interestingly, Bowman found no plateauing effect, with the steady decrease in accuracy continuing through to eight steps. This seems to indicate that attentional control demands do not simply ‘run out’ at some point (causing outright failure), as may be indicated by an attentional ‘capacity’, but rather, the steady increase is reflective of concomitant increases in the duration that attention must be kept controlled. This also means the increase in steps is unlikely to be simply due to a build-up of proactive interference (which may implicate binding capacity demands), because we would expect this proactive interference to become exponentially more detrimental as it accumulates throughout the problem.[14] However, it must be cautioned that Bowman presented items in the Swaps task in order of sequentially increasing steps (i.e., participants started with 1-step, then moved on to 2-step, and so on...). This means an asymptote related to attentional ‘capacity’ limits may have been mitigated by a learning effect (Jensen, 1977), as participants systematically learnt to deal with the steadily increasing demands (this point will become particularly relevant in the Discussion of this chapter).

Nonetheless, it appears that each additional step is more demanding of attentional control than of binding capacity, because attention must be kept active and controlled for longer for each increase in steps. This bodes well for an attentional control view of Gf, since both Bowman (2006) and Stankov (2000) found a linear covariance effect for the number of steps relating to Gf. That is, as the steps increased, so too did the relationship of the task to Gf. Stankov concluded that attentional control was the most likely WM demand linked to Gf. However, there are problems with this conclusion. For Bowman’s (2006) data, as discussed, the number of steps increased sequentially throughout the task (rather than being presented in a random order) and thus, the increase could be more related to a learning aspect (Jensen,

[14] It is possible that the proactive interference demands remain consistent because they are replacing themselves rather than growing but the potential combination of orders invariably increases as the number of steps increase. It takes a minimum of six steps to experience all possible combinations of three letter orders (six possible orderings) but this does not account for repeats, which can occur on every non-adjacent step.

1977) than to an increase in attentional control demands. Stankov (2000), meanwhile, did randomize the order of items, solving that concern for attentional control explanations. However, although the overall linear performance covariance of Steps with Gf was significant, a closer look at the correlations reveals the trend is not as smooth as the performance effect. Rather, there is a large jump going from 1-step (*r* = .248) to 2-step (*r* = .401) but then it quickly plateaus at 3-step (*r* = .429) and 4-step (*r* = .414). Thus, there may be a qualitative difference between 1-step and 2-step unrelated to attentional control occurring. This difference could be as simple as the visual presentation of the ordering for the first step. That is, in the first step, the letters to-be-rearranged are given on the screen. In every subsequent step, the letters to-be-rearranged are ‘presented’ only in the active, direct-access region of WM. Thus, 1-step problems may be qualitatively different in their attentional control demands, not because they require a stepwise, linearly decreasing amount of attentional control; but because they only need to be enacted on a visually presented and accessible arrangement of letters. It is also possible the qualitative difference is an artefact of the ceiling effect occurring on 1-step items. Although the success rate of 1-step items is not at ceiling (90% accuracy), the task does appear quite simple and this deficit from perfect performance may simply indicate a failure to understand the instructions. Regardless, to circumvent the confound of visual presentation,[15] the current experiment included the manipulation of steps with three levels, all requiring active representation of the letter arrangement in direct access: 2-step (2S), 3S, and 4S items. We expect to see a linear decrease in performance associated with the linear increase in steps, because the attentional control demands are prolonged as steps increase. However, we expect to see no covariance of

[15] A pilot test also revealed the 1-step items were virtually unusable in the analyses, with near-perfect performance (M = 98%). Thus, although there was theoretical reason not to consider 1-step items, there was also a practical reason in maximizing the value of participant time.

this linear effect with Gf because research in prior chapters (Chapter III on the LST and Chapter V) have indicated that increasing attentional control demands are not related to Gf.

6.1.2. Letters (and Systematicity)

The second manipulation is (to our knowledge) a novel one for the Swaps task. The traditional implementation of the Swaps task includes a fixed three letter arrangement (e.g., T Q X) with varying number of swap steps (Stankov, 2000). However, as described earlier, this manipulation relates primarily to attentional control, since the actual binding demands of the task remain the same regardless of the step that the respondent is up to: there are always three letters to work with. By increasing the number of letters, we increase the number of constituent elements (letters) involved in each step – the binding capacity required is increased. Although each independent step still only requires the exchange of two letters (two letters are being unbound, exchanged, and bound to new positions), the 'capacity' demands of the task are increased because the full set of letters constitute additional bindings in the direct access region: more letters (contents) and more positions (contexts) need to be worked with throughout the problem. It should be cautioned now that the distance of these swaps and the resulting systematicity does impact on these capacity demands – this section will return to this point shortly. For now, it is simply important to outline that the current experiment includes three levels of letters, including the default level (3-letter/3L) and two additional levels (4L and 5L); reflecting (all else being equal) increases in binding capacity. We expect to see decreases in performance as the number of letters increases, as with other binding increases (e.g., Chapters III-V). Unlike steps (attentional control), we do expect to see covariation of this linear performance effect of Letters with Gf, because the additional bindings increase the capacity demands on the direct access region (Oberauer et al., 2007), which we have observed in earlier chapters. The additional bindings demanded in the *access-random* over *access-fixed* (three to one) in the ACT related to Gf, as did the increase in RC in the LST in 4D items compared to 2D items (though this was not a linear complexity effect because 3D items failed to consistently differentiate from 2D items).

There are two provisos to consider for the Letters manipulation. The first is that the number of letters could be considered raw storage capacity, rather than binding capacity. As detailed throughout the thesis, the current perspective is that the direct access region is limited through limits on binding capacity, rather than the more ambiguous ‘storage capacity’. Through earlier studies in this thesis, we concluded that ‘storage’ capacity could refer to both active and passive elements (i.e., bound within direct access, and passively activated but outside direct access), but passive elements did not seem to relate to Gf (Bateman, 2015; Chapter IV). Binding capacity more specifically refers only to active relations in the direct access region. The next concern then, is whether it is possible that some letters within a letter set of a Swaps problem could be considered ‘passive’. This is related to the second proviso.

The second proviso is more complex, to the point it requires an additional experimental manipulation. As we have identified in earlier chapters, problems of high complexity can be systematically chunked down, depending on how systematic the elements are. This is particularly important to a manipulation of Letters. In the basic Swaps task, there are only three letters. All three letters must be used in problems containing two or more steps, because if only two letters are used, then consecutive steps would simply be repeating (or reversing) the prior step, cancelling both steps out. For instance, in the item [T Q X | Swap 1 with 2 | Swap 2 with 1], the answer is simply [T Q X] because the two swaps used the same two letters. Thus, all three letters must be used within two consecutive steps to ensure this does not occur. This same restraint does not apply to items with more than three letters, because the three-letter logic that ensures steps are not repeated can be applied to four letters, effectively leaving the fourth letter out of all the steps. This also applies to five letters, except that now two of the five letters can be excluded from the instructions. This has important implications for our interpretation of the Letters manipulation as being differentially demanding of binding capacity because, as we know from the ACT (Chapter IV), the systematic reduction of capacity demands can lead a three-binding problem to be completed as a one-binding problem. In this case, a five-letter problem can be solved as a three-letter problem with the simple application of systematicity. For instance, consider the item [T Q X B L | Swap 1 with 2 | Swap 3 with 2 | Swap 1 with 3]. In this item, the two adjacent letters of [B L] can be kept systematically fixed throughout the problem. The item can be solved as a 3-letter item with [T Q X] becoming [X T Q], then the [B L] can simply be addended to the response after the steps have been resolved as [X T Q B L]. Although the 4L and 5L conditions are novel, prior evidence from the Swaps task demonstrates the importance of considering the position of Letters. Unpublished data collected by Birney (n.d.-a) indicates that the distance of the swap determines the likelihood of the error: a distant swap (*Swap 1 with 3*) leads to higher error rates than a close swap (*Swap 1 with 2*). Our findings from the ACT study in Chapter IV indicate this could be another example of the impact of systematicity (Halford et al., 1998) where the isolated digit in the close swap (e.g., the *'3'* in *Swap 1 with 2*) reduces the binding demand because it can be held systematically fixed during the unbinding/rebinding process (like our *access-fixed* condition). Although the current study does not specifically consider the positioning, this Birney (n.d.-a) data nonetheless demonstrates the important of considering how the increase in the number of letters influences more than just the overall binding capacity demands.

Although we can record the steps of each item, the nature of increasing letters means that items of higher letters and fewer steps have a higher chance of incidental systematicity. Because the Swaps problems would be randomly generated, this may result in inadvertent bias in items of higher letters on aggregation of the conditions just due to the increased number of permutations that exclude one or two letters. Thus, instead of letting systematicity occur naturally, we experimentally manipulated the presence of systematicity as a third manipulation (Steps, Letters, Systematicity). *Systematic* items were generated using 3L generation logic only, effectively reducing the binding demands of all systematic items to 3L. Specifically, 4L systematic items would have one letter fixed in position (not used in any swap steps) while 5L systematic items would have two adjacent letters (a bigram) fixed in position (not used in any of the swap steps). *Non-systematic* items were generated with code that ensured all letters were used where possible. Because 3L items must use all letters in the swaps (otherwise they simply reverse the same swap repeatedly), 3L items would not be included in analyses of systematicity (i.e., 3L items are generated identically in the systematicity 'on' and systematicity 'off' conditions). Although we do not have any theoretical reason to suspect that 4L and 5L should differ (since both rely on 3L logic), the fixedness of one letter as opposed to a bigram may still cause some differences. Thus, we first consider 4L and 5L items independently for the purpose of verifying their equivalence in the further investigation of systematicity. We expect trials with systematicity *on* to be easier than those with it *off*. The increase in accuracy for having systematicity on for 4L and 5L items should end up with performance similar to that seen in 3L items, minus any deficit related to the additional storage of a single letter (4L) or bigram (5L), which we do not anticipate being substantial. There may however, be an interaction with Steps, as the benefit of systematicity is amplified for items with higher steps, because the fixed letters are held in place for longer. Given the results of the ACT, we would expect participants higher in Gf to be more capable of dealing with non-systematic items, where the binding capacity demands are highest.

6.1.3. Hypotheses

The current experiment was designed to further specify the demands associated with binding capacity and attentional control in WM. Specifically, we aimed to uncouple the frequently seen overlap between binding capacity and attentional control. In aim of this goal, we also consider the important role of systematicity, which we identified in Chapter IV as an important determinant relating to the true binding capacity demands of the task. As with studies in prior chapters, the approach was to experimentally manipulate the core task (the Swaps task) to differ in theoretical demands and observe how these experimental manipulations changed the variance in prototypical tasks representing key constructs such as Gf. Unlike prior experiments the manipulations (Letters and Steps) in the Swaps task are expected to be quantitative in nature.[16] Thus, rather than using each condition as separate predictors in a linear regression, we use an analysis of covariance (ANCOVA) for each hypothesis, with the expectation that the Letters and Steps manipulations can be plotted as linear functions. We then compare how latent variables of *Gf* and *WM* covary with these linear functions.

For Steps (attentional control), we hypothesized a linear decrease in performance associated with increases in the number of steps (2S > 3S > 4S), as seen in Stankov (2000). However, unlike Stankov (2000), we do not predict that steps will covary with Gf, because the lack of 1-step items removes the qualitative difference that occurs as a result of arranging letters visually presented (as opposed to those only in the direct-access region). In other words, although attentional control demands increase consistently and linearly with increases in Steps (as they require longer attentional control), these demands are hypothesized to be *not* related to Gf. Thus, increased steps will *not* increase the relationship to Gf. Similarly, Bowman's (2002) sequential ordering of Steps means the linear covariance with Gf observed by him may simply be due to learning (Jensen, 1977). However, attentional control demands

[16] RC in the LST is theorized to be, and was expected to be, a quantitative manipulation also, with the pattern of 2D > 3D > 4D. However, the analyses consistently revealed a pattern more like 2D = 3D > 4D, and the task breakdown provided in discussions indicate a more qualitative difference between 4D items and the other two RC levels. For the RMT, the *same*, *different*, and *ascending* conditions were never theorized to be a continuous scale, only that the binding demands in *different* were theorized to be higher than those in *same* and *ascending*, though *same* and *ascending* still differed in the visual similarity of the elements to-be-integrated.

are more likely to be tapped through traditional WM tasks such as complex spans (used often by, e.g., Engle, 2018), so we do expect this linear function of Steps to covary with our *WM* factor.

For Letters (capacity), we hypothesized a linear decrease in performance associated with increases in the number of letters involved in the problems (3L > 4L > 5L). Although this is a novel manipulation, the prediction comes from increased binding load on the direct access region. Previous chapters have observed a decrease in performance related to increased binding demands (access-random vs. access-fixed in the ACT; different vs. same in the RMT).

Unlike Steps, we do expect to see a covarying effect of Gf, such that increases in letters are associated with concomitant increases in the relationship to Gf. This covariation has been somewhat seen in earlier chapters. Although this has not been consistently demonstrated, various reasons have been explored for why this is not showing consistently, including that Gf tasks may not necessarily differ in binding capacity demands, even though they do, more generally, tap a fundamental binding function. The overall implication of binding theory is that relational integration demands acts upon both WM and Gf (the relational integration hypothesis). Although these demands are assumed to be through the number of active bindings in the direct access region (Oberauer et al., 2007), it is not yet clear whether this can be most appropriately observed through linear increases in binding 'capacity' demands (as Chapter V failed to find this difference between same vs. different). In the current experiment, we assume they are (Oberauer et al., 2007) and expect to see the linear effect of Letters leading to concomitant increases in covariation with both Gf and WM.

Finally, for Systematicity (fixedness of letters and bigrams throughout 4L and 5L problems, respectively), we predict that systematic items will lead to higher performance than non-systematic items, and this effect will be amplified for items benefitting the most from systematicity, which is those of higher steps. This performance increase should put systematic 4L and 5L items on the same performance level as 3L items, minus any deficit associated with carrying the fixed letters across the problem steps or reintegrating them with the final step (these demands are anticipated to be minimal). In line with the results from the ACT (Chapter IV), we also hypothesize that items with systematicity OFF will correlate more with Gf than items with systematicity ON, because they increase the binding demands of the problem (in line with the prior ‘capacity’ hypothesis). Specifically, systematic 4L and 5L items should correlate with Gf similarly to 3L items, while non-systematic 4L and 5L items will correlate more with Gf in correspondence with the increasing binding capacity demands.

6.2. Method

6.2.1. Participants

There were 106 participants who participated in exchange for course credit. The mean age was 19.90 (SD = 3.85) and there were 74 females (69.8%). This is the same dataset from Experiment 3 of Chapter III on the LST, though the focus of the analyses in that chapter are on the LST rather than the Swaps task.

6.2.2. Measures

Swaps Task

Participants completed 36 items of the Swaps Task as described in Stankov (2000). On each item, participants were presented a problem page with a set of letters (e.g., B K M) arranged towards the top of the screen and below, several lines of instructions (steps) that instructed the participants to swap the order of letters (e.g., “Swap 1 with 2 | Swap 1 with 3”).

The number of *Letters* varied between three to five. The number of *Steps* varied between two to four. Letter arrangements were randomly generated by selecting from any of the consonants of the English alphabet with the only constraint that all letters in a problem had to be unique. Swap steps were generated randomly with one constraint: two consecutive swaps could never cancel each other out (i.e., “Swap 1 with 3” could never be followed by “Swap 1 with 3” *or* “Swap 3 with 1”). This constraint meant that step generation would naturally prioritise using letter positions that had not previously been used. There were 12 of each number of letters (3L/4L/5L) and 12 of each number of swaps (2S/3S/4S) generated to make up the 36 items for each participant, mixed evenly across the two variables (i.e., there were four 3L2S items, four 3L3S items, and so on) and presented in a random order.

The final manipulation was *systematicity*. When systematicity was *off*, the items were generated randomly using the above logic. When systematicity was *on*, the same logic was applied except that 4L and 5L items used the steps generation of 3L items to produce the instructions, such that only three of the four/five letters were actually used in the problem. An additional constraint ensured that the one/two excluded letters were chosen randomly from all available letters rather than always excluding the far-right letters. For 5L items, the two excluded letters were always adjacent, making it an excluded bigram. Systematicity was applied evenly, such that half of each type of item (e.g., half of 4L, half of 2S) had systematicity off and half had systematicity on (for 3L items, systematicity being on does nothing because it uses the same 3L logic either way).

When participants were ready to respond, they pressed spacebar which progressed the program to a response page that was blank apart from a textbox where they could type and submit their response. This is different from previous research (e.g., Bowman, 2006) which presented the possible response options (all possible orderings) to choose from. The current method was preferred to prevent guessing and because it would be unpractical to display the 120 possible combinations of letters for the novel 5L items (as opposed to only six combinations in 3L items). Feedback was given on each item along with the answer.

Gf Tasks

The same 20-item version of Raven's APM from earlier chapters was included. The same Letter Series from earlier chapters was included. The LST from Chapter III (Experiment 3) was also administered but the data is not included in the analyses presented here, primarily because the lack of rule induction makes it unsuitable as a Gf task (as described in Chapter II).

WM Tasks

The same spatial *n*-back, Operation Span (OSPAN) and Symmetry Span (SSPAN) from earlier chapters were included. For both OSPAN and SSPAN, the dependent variable was total number of elements (letters/squares) recalled across the task and the processing cut-off was set at 80% accuracy. Scores below this threshold were removed and set as missing data.

6.2.3. Procedure

Participants were tested in groups of one to ten in computer labs at the University of Sydney. The testing sessions were 90-minutes long and participants were instructed to complete as many of the seven tasks as possible in the 90 minutes. The tasks were presented in a random order, except for SSPAN which was always presented last. This was because the SSPAN was thought to be the most disposable in the analyses, considering the OSPAN was already included. In total, 80 of the 106 participants completed all tasks including the SSPAN, while a small amount (two to four participants) completed five of the seven tasks, missing one other task. The implications of this missing data on the SSPAN are described in the Results.

6.3. Results

6.3.1. Descriptives and performance effects

Descriptives for all tasks are presented in Table 6.1. As can be seen among the Swaps descriptives, performance was considerably better than chance in all Letter conditions, despite chance level changing across the manipulation (i.e., 1 in 6 (16.7%) for 3L, compared to 1 in 120 (0.8%) for 5L[17]). Means and standard deviations were generally as expected across the tasks, with similar distributions to those seen in the prior studies. It is important here to once again point out the difference between OSPAN and SSPAN that is likely a symptom of the choice to set the SSPAN as always last in the task order (as opposed to randomized with the other tasks). The implications of this task ordering, and related subsample issues are described in detail in Section 3.9.2 (which used the same dataset). For consistency and due to concerns over selection bias, the same approach to analyses has been used in this study. That is, SSPAN is excluded from the majority of analyses to ensure the full dataset (n = 106) is employed. Where possible, data on the SSPAN is still mentioned, but findings on this subset should be interpreted cautiously.

[17] This chance level calculation assumes that participants know what letters were involved in each problem. Actual chance level may be somewhat lower than this, considering the letters were randomly generated throughout the problem.

Table 6.1. *Descriptives for each Swaps condition (averaged over other variables), and the criterion measures for Gf (APM, Letter Series) and WM (*n-*back, SSPAN, OSPAN). The* 'n' *column refers to participants who completed the task (max is 106).*

Task	Condition	Mean	SD	*n*
Swaps	**Total (proportion correct)**	**.62**	**.19**	**106**
	3-Letter (3L)	.72	.19	106
	4-Letter (4L)	.61	.22	106
	5-Letter (5L)	.52	.25	106
	2-Step (2S)	.77	.17	106
	3-Step (3S)	.62	.25	106
	4-Step (4S)	.47	.25	106
	Systematicity OFF *(4L+5L only)*	.55	.23	106
	Systematicity ON *(4L+5L only)*	.58	.24	106
Raven's APM	**Total (proportion correct)**	**.63**	**.18**	**104**
Letter Series	**Total (proportion correct)**	**.68**	**.15**	**105**
OSPAN Letters	**Total (proportion recalled)**	**.90**	**.13**	**101**
SSPAN Letters	**Total (proportion recalled)**	**.79**	**.16**	**80**
***n*-back**	**DV***	**2.63**	**1.68**	**105**

**n*-back mean is reported as dependent variable (hits minus false alarms) rather than proportion of items correct.

As seen in Figure 6.1, performance was affected by both letters and steps, with linear decreases in performance as either variable went up. A repeated-measures ANOVA was performed to test these performance effects statistically. The linear effect of letters was indeed significant, $F_{1,104} = 105.54$, $p < .001$, $\eta_p^2 = .504$, such that as letters increased, performance decreased. The linear effect of steps was also significant, $F_{1,104} = 158.60$, $p < .001$, $\eta_p^2 = .604$, such that as steps increased, performance decreased. There was also a significant interaction between these two linear effects, $F_{1,104} = 5.96$, $p = .016$, $\eta_p^2 = .054$, such that the increase in steps led to sharper decreases in accuracy for higher letter conditions (see Figure 6.1). Neither quadratic effect was significant ($F_{1,104} = 2.27$, $p = .135$, $\eta_p^2 = .021$ for Letters; $F_{1,104} = .018$, $p = .895$, $\eta_p^2 < .001$ for Steps).

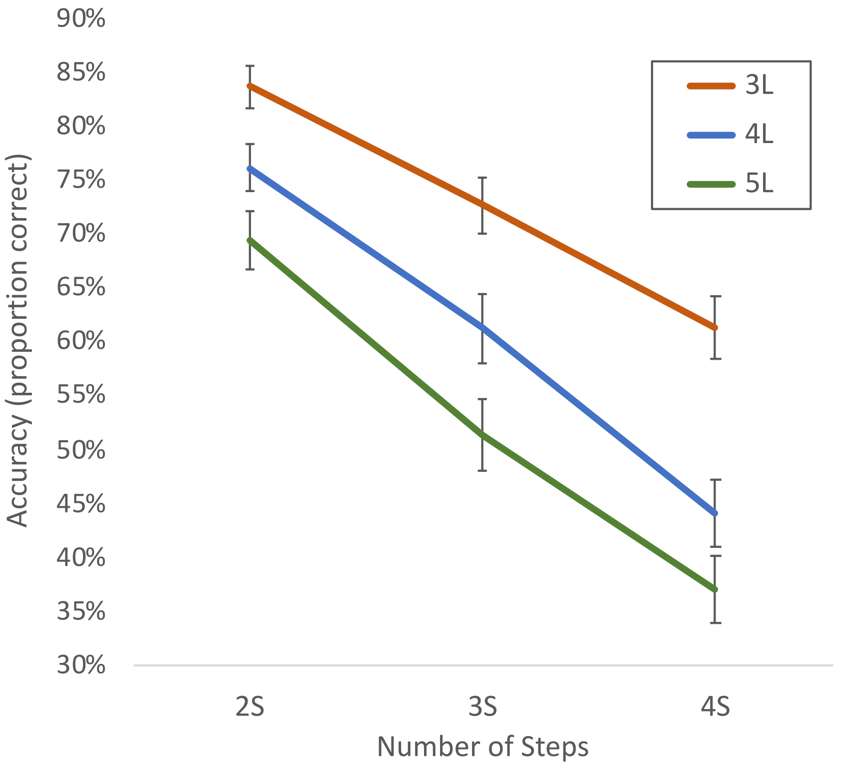


Figure 6.1. Accuracy (proportion correct) split by Steps and Letters. Error bars represented standard error. Line graphs are plotted because both variables are theoretically linear.

To determine the influence of systematicity, a repeated-measures ANOVA was run again, this time including systematicity as a third variable. 3L items were removed from the Letters variable (because they are unaffected by systematicity[18]), meaning letters varied only between 4L and 5L. Overall, systematicity did not affect performance, $F_{1,105} = 3.38$, $p = .069$, $\eta_p^2 = .031$. However, as seen in Figure 6.2, there was a significant interaction between systematicity and the linear function of steps, $F_{1,105} = 15.73$, $p < .001$, $\eta_p^2 = .130$, such that systematicity made items of higher steps disproportionately easier than items of lower steps. There was also a significant interaction between systematicity and the quadratic function of steps, $F_{1,105} = 4.83$, $p = .030$, $\eta_p^2 = .044$. As seen in Figure 6.2, these linear and quadratic interactions are represented by the sharp decrease in performance going from 2S to 3S and

[18] There was indeed no difference between systematicity ON (M = 1.40) and systematicity OFF (M = 1.46) for 3L items, $F_{1,105} = 0.30$, $p = .587$, $\eta_p^2 = .003$.

then the more subtle decrease from 3S to 4S when systematicity is *on*, which is not replicated when systematicity is *off*. The interaction between letters and systematicity was not significant, $F_{1,105} = 0.04$, $p = .851$, $\eta_p^2 < .001$, nor was the three-way interaction with steps included, $F_{1,105} = 1.76$, $p = .187$, $\eta_p^2 = .017$. However, in comparing the 3L line in Figure 6.1 to the systematic 4L and 5L lines in the right of Figure 6.2, it is evident that systematicity did not, in fact, make the 4L and 5L items equivalent to 3L items in difficulty. A repeated-measures ANOVA comparing Letters (3L vs. a composite of systematic 4L and systematic 5L items) and the linear effect of Steps (2S vs. 3S vs. 4S) confirmed this was the case, with 3L items having significantly higher accuracy than the systematic composite, $F_{1,105} = 57.88$, $p < .001$, $\eta_p^2 = .355$; and this difference was not influenced by an interaction with Steps, $F_{1,105} = 0.14$, $p = .714$, $\eta_p^2 = .001$.

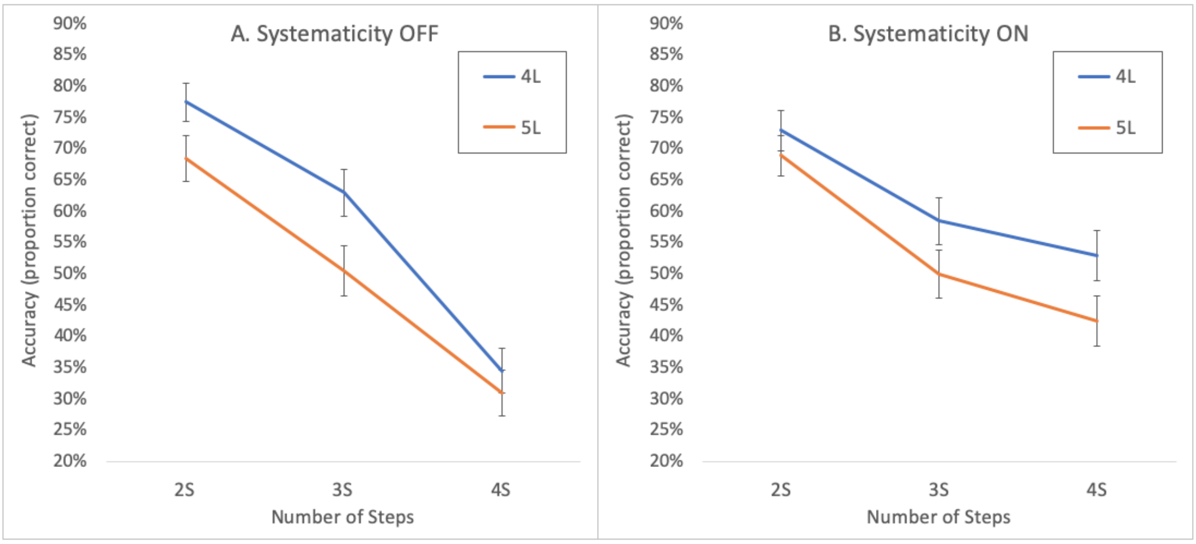


Figure 6.2. Accuracy (proportion correct) split by steps, letters, and systematicity. Error bars represented standard error. 3L items omitted as they are unaffected by systematicity.

6.3.2. Covariation with WM and Gf factors

Given that both letters and steps influenced performance, the next step was to determine whether these linear functions (and the systematicity*steps interaction) were moderated by performance on the Gf or WM tasks. A *Gf* factor was created with principal axis factoring with varimax rotation on the two Gf measures: APM and Letter Series.

Together, the extracted factor accounted for 71% of the variance in the two measures. Similarly, a *WM* factor was extracted (with the same method) on *n*-back and OSPAN (SSPAN excluded for reasons detailed above). The *WM* factor accounted for 66% of the variance in the two measures.

Task-level correlations are presented in Table 6.2. To test the moderating effects of Gf and WM, two separate ANCOVAs were run replicating the initial ANOVAs on Letters and Steps with either *Gf* or *WM* added as a covariate. *Gf* was a significant moderator of the linear function of Letters on performance, $F_{1,101} = 7.69$, $p = .007$, $\eta_p^2 = .071$, suggesting influence of Gf on the ability to deal with additional letters, thought to represent binding capacity demands. *Gf* was not a significant moderator of the linear function of Steps, $F_{1,101} = 2.37$, $p = .127$, $\eta_p^2 = .007$, suggesting that Gf has no influence on the attentional control aspects of the task. Neither quadratic functions were significantly influenced by Gf, nor was there a three-way interaction between Letters, Steps, and *Gf*, all *p*'s > .05.

Table 6.2. *Correlations for task measures.*

	Gf	APM	L-Series	*WM*	n-back	OSPAN	SSPAN
Swaps	**.58**	**.52**	**.46**	**.58**	**.52**	**.35**	**.31**
Gf	-	**.84**	**.84**	**.49**	**.52**	**.20**	**.55**
APM		-	**.42**	**.41**	**.48**	.15	**.57**
L-Series			-	**.40**	**.40**	.19	**.32**
WM				-	**.81**	**.81**	**.43**
n-back					-	**.32**	**.46**
OSPAN						-	**.22**
SSPAN							-

N=106; bold coefficients p < .05. *Gf* is the factor extracted from APM and L-Series. *WM* is the factor extracted from *n*-back and OSPAN.

Repeating the analysis with *WM* as a covariate resulted in a significant moderating effect of *WM* on the linear function of Letters, $F_{1,97} = 6.46$, $p = .013$, $\eta_p^2 = .062$. *WM* did not moderate the linear function of Steps, $F_{1,97} = 0.025$, $p = .875$, $\eta_p^2 < .001$; though *WM* did moderate a positive quadratic function of Steps, $F_{1,97} = 8.76$, $p = .004$, $\eta_p^2 = .083$, such that

3S items were most strongly influenced by *WM*. There was no three-way interaction between Letters, Steps, and *WM*, $F_{1,97} = 0.744$, $p = .390$, $\eta_p^2 = .008$. The two hypothesized relationships (*Gf* by Letters and *WM* by Steps) are plotted in Figure 6.3, demonstrating the present but inconsistent linear covariation of Gf with Letters. More specific condition-level breakdowns are provided in Table 6.3, which demonstrates how the relationship to Gf changes little with differences in Steps but does generally increase with Letters.

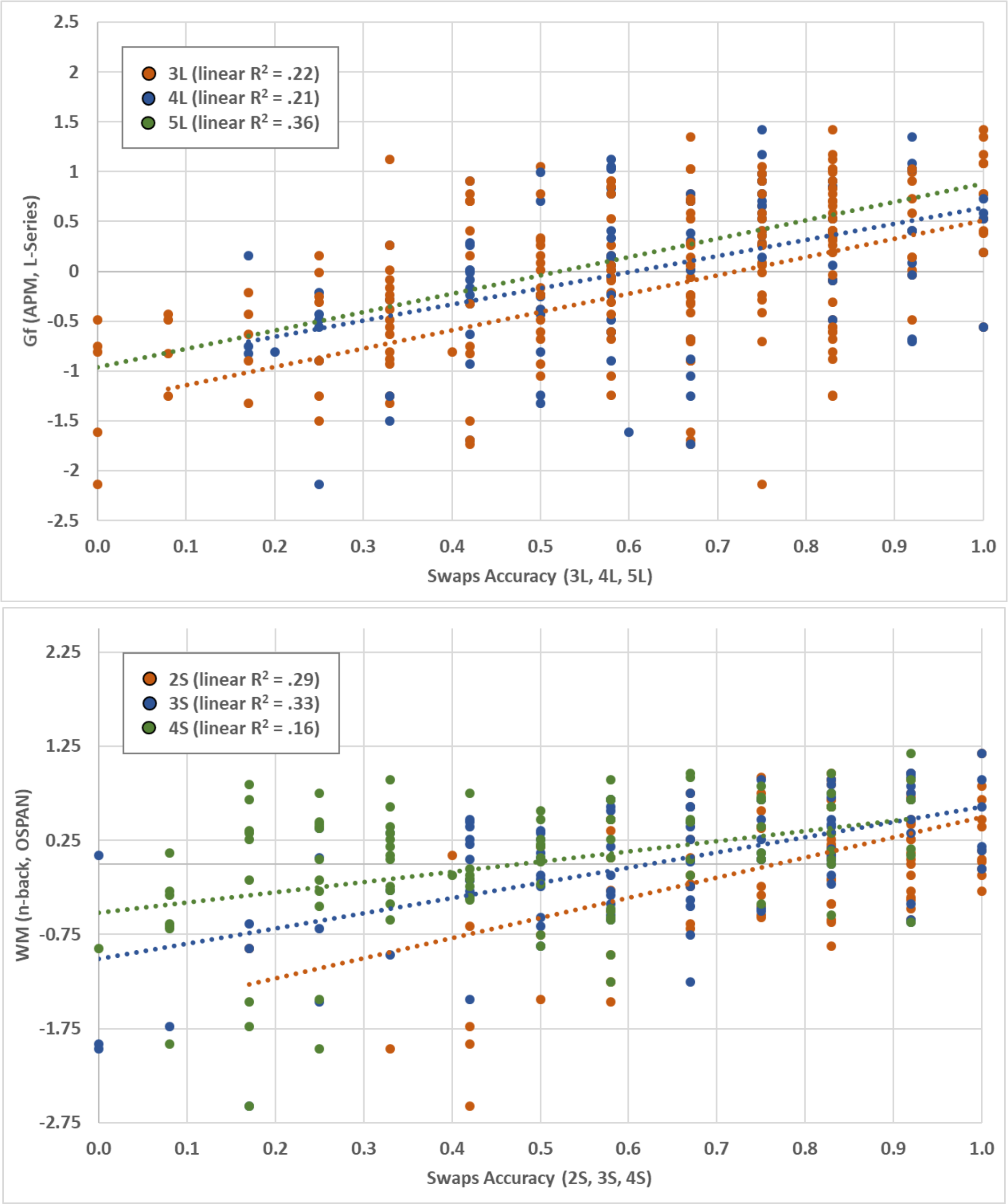


Figure 6.3. Scatter plots of the two hypothesized covariation effects: that (top) Letters would covary with Gf; and (bottom) Steps would covary with WM. For the top graph, the trendlines demonstrate that the hypothesized covariation of Letters with Gf is present (primarily through the high correlation between 5L and Gf) but weak (because of the weaker correlation between 4L and Gf). The bottom graph demonstrates that the covariation of Steps with WM fails because the 4S items have the weakest linear correlation to WM.

Table 6.3. *Condition-level squared correlations (predictor in bold at top-left cell of each inset table)*

r^2 ***Gf***	2S	3S	4S	*Total*	r^2 ***WM***	2S	3S	4S	*Total*
3L	.15	.08	.12	*.22*	3L	.10	.21	.07	*.21*
4L	.06	.16	.13	*.21*	4L	.14	.23	.10	*.27*
5L	.21	.22	.22	*.36*	5L	.25	.22	.13	*.30*
Total	*.22*	*.23*	*.24*	*.32*	*Total*	*.29*	*.33*	*.16*	*.35*

N=106.

Finally, for systematicity, overall the correlation between *Gf* and the Swaps task was higher when considering *systematic* items (r = .552) compared to *non-systematic* items (r = .469). However, both of these correlations were significant, and the two correlations did not significantly differ from one another, z = -1.19, p = .117. Based on the results of the accuracy ANOVA (which found an interaction between systematicity and steps (both linear and quadratic), but not between systematicity and letters), closer analyses on systematicity were considered by excluding the effects of Letters (4L and 5L items were grouped and 3L were excluded altogether, such that the dependent variable was averaged across 4L and 5L items) and comparing different levels of Steps. As a reminder, the mean scores between the Steps levels with Systematicity ON and OFF can be seen in Figure 6.2, which demonstrates a marked increase in performance at 4S items with systematicity ON, but not at lower steps. However, in an ANCOVA with Steps and Systematicity as within-subject effects and *Gf* as a covariate, *Gf* did not significantly moderate systematicity overall, $F_{1,101}$ = 0.366, p = .546, η_p^2 = .004; nor did it moderate either the linear ($F_{1,101}$ = 2.091, p = .151, η_p^2 = .020) or quadratic functions ($F_{1,101}$ = 1.754, p = .188, η_p^2 = .017) of Steps. None of the interactions between Steps, systematicity, and *Gf* were significant, p's > .05. Thus, overall, systematicity did not seem to influence the relationship with Gf, even at its most impactful level (4S).

6.4. Discussion

The study presented in Chapter VI was aimed at demarcating attentional control and binding capacity demands in the Swaps task. Although earlier chapters have made a compelling case for the importance of binding capacity as in relational integration, it has been difficult to fully demarcate relational integration explanations of the results from attentional control. The Swaps task was chosen as an ideal measure, as it allowed us to systematically manipulate attentional control demands separately from binding capacity by manipulating the number of steps and the number of letters in each problem, respectively. While the actual binding demands remained the same between steps, the attentional control demands increased simply because the active elements had to be kept in their right place for longer with more steps. Binding capacity meanwhile, was manipulated through the number of elements involved in each problem (the number of letters). Thus, an item could be low in attentional control demands (with fewer steps) but high in capacity demands (with more elements involved). Chunking was also accounted for by considering *systematic* and *non-systematic* problems, which kept some elements in each problem systematically fixed, effectively mitigating the capacity demands to match items of lower letters.

The overall goal of distinguishing attentional control from binding capacity was successful. Increases in steps (attentional control) and letters (capacity) both produced steady and substantial increases in difficulty (refer to Figure 6.1). Although there was an interaction between the two variables, this is not surprising given the difficulty in uncoupling attentional control from capacity. Overall, the difficulty range of the Swaps task with these two manipulations was ideal. That is, the easiest items sat comfortably below ceiling (M = 84%) and the hardest items sat comfortably above the floor (M = 36%).

In terms of cognitive correlates, we compared both a *WM* and a *Gf* factor to these linear performance effects. For the *WM* factor, it was hypothesized that both Letters and

Steps would relate to WM (but only Letters to Gf). The results indicated that there was Letters by WM covariation, but not Steps by WM. This puts some doubt over the attentional control demands represented by the increasing Steps, since prolonged attention is required in both increases in Steps and increases in the set size of our criterion tasks, operation span and *n*-back. It is possible that the way the steps are quantified do not match evenly to the increasing demands of our criterion tasks, OSPAN and *n*-back. As illustrated in the introduction, the increasing steps do not widen the scope of attention control, they only ensure that attentional control on the same select number of elements must be maintained for longer – which is what Kane et al. (2001) claims to be a defining characteristic of attentional control. The increase in difficulty is not related to changes in the number of items, only how long it must be maintained. In this way, it differs from OSPAN where *both* the duration and the number of elements increases as the set size of each trial goes up. While the *n*-back does keep the element size consistent (at *n*), considerable, incremental demands are also incurred by having to drop (unbind) the memory element *n* - 1 steps back. This can be particularly difficult when the element is several steps back, where the memory system is reliant on an ongoing chunk being formed. In the Swaps task, only the prior step is being dropped as the memory system remains active on the current step. Thus, the fairly crude prediction that Steps would covary with traditional WM measures simply came about through theoretical convenience though we see, once again, the problems with taking a latent variable analysis in favour of a more thorough task analysis.

With that said, the results for Gf were more convincing. Overall, there were strong correlations between Swaps and our Gf measures, APM (r = .52), Letter Series (r = .43), and the latent *Gf* variable (r = .58). However, our main goal (as always) is to relate the separate manipulations of the primary task (Swaps) to Gf in line with the hypotheses. On one hand, the linear effect of Steps did not covary with Gf: all three Steps levels resulted in similar correlations to the latent variable of *Gf*. If increased attentional control demands are related to Gf (e.g., Kane et al., 2001), the increased attentional control demands in the higher Steps levels should have drawn on more Gf resources, but this was clearly not the case. On the other hand, there was a more positive result for binding capacity views, as Gf did significantly moderate the relationship between Letters and performance, such that increased Letters led to a higher relationship with Gf. This result needs some caution though as, like what was has been seen before (in the LST), this effect was not a result of a clear linear pattern. Rather, a sudden jump in the correlation occurred for 5-letter items, while 3-letter and 4-letter items had similar correlations to Gf. This monotonic effect was the same criticism leveraged against Stankov's (2000) attentional control explanation in the introduction. Thus, there may be some qualitative difference occurring at 5-letter items.

Although the incorporation of systematicity and the binary swaps occurring in the swaps task (i.e., each step only involves two elements changing regardless of the number of letters) makes it somewhat difficult to align the task to capacity theories such as Cowan (2001) or Halford et al. (1998), it is nonetheless intriguing that the qualitative difference occurs at 5-letters, when in both Cowan and Halford et al.'s theories, four elements is the 'magical' number that represents the upper limit of capacity. This difference between four and five letters is thus worth considering in more depth, as it may represent the difference between items that fall within a natural capacity against those that exceed capacity limits.

A capacity limit of four does provide more context to our findings on *systematicity*. As theorized in the introduction, and discussed in-depth throughout this thesis, findings on binding capacity cannot be concluded without also considering the role of systematicity: the ease at which a set of bindings can be systematically reduced to fewer bindings. In the standard 3-letter Swaps task, systematicity is not relevant. There are only three elements in the problem and every step involves exactly two elements. Because every second step involves an element combination that *must* be different from those in the prior step, every element is used within one step of its prior usage. Thus, all three letters are used on every consecutive step, demanding exactly three bindings for the entirety of the problem. This changes with additional letters, because the number of elements in the problem goes up but the number of elements in each step remains at exactly two. Thus, using 3-letter swap logic, a 4-letter item could forgo one of the elements entirely in all the steps of the problem (and a 5-letter item can forgo up to two of its elements entirely). Thus, to fully consider increases in letters, systematicity had to be accounted for. Rather than leaving it to random item generation (which would bias systematicity towards higher steps and higher letters), we introduced a systematicity variable to contrast the effects of *systematic* and *non-systematic* items, to determine the impact of systematicity on letters and steps. It was hypothesized that systematicity would effectively turn 4L and 5L items into 3L items (in terms of difficulty). This is because systematic 4L and 5L items can be solved using a 3L approach, with the fixed letters addended afterwards when the response is actually entered. The actual results were not this simple. Systematic 4L and 5L items were still considerably more difficult than 3L items, and the benefit of systematic 4L and 5L items (over non-systematic 4L and 5L items) was mostly marginal *except* for 4S items. At 4S, the benefits of systematicity were more clearly apparent (see Figure 6.2). As it turns out, 4S items also represent the items where the emerging systematicity is more clearly apparent to participants. At 2S for instance, the benefit of systematicity is low (because elements are only held over two steps) and may be easily offset by an unexpected ‘usage’ benefit as a result of actually *using* the element, as opposed to not using it. This could well be the case, considering that recall for the ABC variables in the ACT study was superior when the ABC variables were actually used as part of the problem (access) compared to when they were not (retention). This conundrum between systematicity benefits and access benefits in low-step items could possibly be answered by a more in-depth response analysis. For instance, if erroneous responses tend to contain only elements that were actually used, it would indicate a benefit for usage that is not seen in fixed letters.

These unexpected (but potentially explainable) results described above make it quite difficult to come to interpretations about the overall point of the systematicity variable, which was to see if the linear effect of Letters (capacity) is diminished or enhanced by systematicity. It is however, particularly important to fully dissect the systematicity effect, since (as described above) an initial hypothesis was that only high Gf participants would be able to deal with the higher binding demands involved in high-letter items, provided that all participants were being facilitated by systematicity. This hypothesis was due to the results of the ACT, where restricting systematicity (necessitating higher binding demands) led to the task increasing its correlation with Gf. We did not find this same effect in the Swaps task, as both systematic and non-systematic items correlated with Gf. If anything, the effect was reversed because the systematic items had a higher correlation than the non-systematic items. Despite this reversal, these results do not completely contradict the ACT findings. For a start, the only manipulation in the ACT was the ABC mappings (i.e., the only independent variable was how the ABC set was involved in the arithmetic) while the Swaps had other manipulations (steps and letters) that coincided with the systematicity manipulation. The downside of this in the ACT was that we could not demarcate binding from attentional control demands, but it did make the operational outcome of the task cleaner than what was observed in the Swaps task.

The possible trend towards a systematicity by Gf interaction observed in the Swaps task indicates that there may be more strategic exploitation of the systematicity occurring. That is, in the ACT, everyone improved as a result of the systematic bindings (i.e., everyone is taking advantage of systematicity) but *only* high Gf participants could deal with the additional bindings in non-systematic trials. The active retention of the ABC variables in the ACT was pivotal to item success because once they were lost, the item was doomed. Being presented with a systematic, fixed ordering (ABC=XYZ) made this active retention immediately and noticeably more comfortable. As seen in the distribution of costs (Figure 4.5d, in Chapter IV), everybody was taking advantage of systematicity. Where individual differences were more often seen (the flatter, wider cost distribution in Figure 4.5e) was in the cases where systematicity could not be used. This is where capacity was stretched, and a truer proxy of Gf was obtained. Compare this to the Swaps task. It is possible that low-Gf participants were not even identifying the benefits of systematicity, because it was not immediately obvious that it would confer a benefit. For instance, if the letter sequence [B L K D J] is presented with [K D] systematically fixed, lower-Gf participants may still attempt each swap step with the full sequence [B L K D J], incurring maximum capacity cost. Higher-Gf participants meanwhile, may appraise the sequence of steps and recognise the systematically fixed bigram. This would allow them to work through each step with the fixed bigram excluded, [B L – J]. After the final step, they can then re-insert the bigram. A strategy like this has some minor additional costs (appraising the item, removing the bigram, re-inserting the bigram) but the benefit is a working binding demand of only three elements. Because these additional costs are applied regardless of the number of steps, this strategy may not even be worth doing at low steps but gains purely profitable benefits for each additional step in the item. Of course, this is speculative, given that we cannot tell whether high-Gf participants were actually ‘appraising’ the item. It may instead be possible that high-Gf participants are better or quicker at learning to take advantage of systematicity over the course of the task. This speculation could be confirmed with an item-based analysis that considers the trajectory of performance through item ordering; or with an experimental manipulation that adds a condition where only one step is visible at a time (restricting the ability to initially appraise the step sequence in full). Nonetheless, it is still important to (where possible) consider the difference between high-Gf and low-Gf participants in respect to systematicity, on aggregate.

Given the importance of this test in resolving the contrasting findings of the ACT with the Swaps, a post-hoc analysis was performed to determine how many participants were actually taking advantage of systematicity. The participant pool was mean-split (high Gf > 0.00 on the *Gf* factor; low Gf <= 0.00) to produce two independent samples: high-Gf and low-Gf participants. Their relative performance on each type of item was then considered and the results are presented in Table 6.4.

Table 6.4. *Comparison of systematicity advantage for High-Gf participants compared to Low-Gf participants (with cells critical to the hypothesis highlighted in blue).*

Low Gf **(n = 51)**	3L items	4L+5L Sys OFF items	4L+5L Sys ON items	Systematicity Advantage	Shortfall to 3L items
2-Step items	78%	62%	65%	**+3%**	**-13%**
3-Step items	64%	48%	41%	**-7%**	**-23%**
4-Step items	52%	26%	33%	**+7%**	**-19%**
Total	64%	46%	46%	**0%**	**-20%**
High Gf **(n = 55)**	3L items	4L+5L Sys OFF items	4L+5L Sys ON items	Systematicity Advantage	Shortfall to 3L items
2-Step items	89%	85%	79%	**-6%**	**-10%**
3-Step items	81%	66%	67%	**+1%**	**-14%**
4-Step items	70%	40%	61%	**+21%**	**-9%**
Total	80%	64%	69%	**+5%**	**-11%**

In line with the theory, high Gf participants overall, gain an advantage from systematicity (+5%) where low Gf participants do not (0%). In particular, high Gf participants gain a considerable systematicity advantage in 4-step items (+21%) – items which are theorized to benefit the most from systematicity. There is even a slight disadvantage to 2-step items for high Gf participants (-6%), in line with the suggestion that there is an additional cost to undertaking a strategy that exploits systematicity (in this case, with no benefit or even a slight cost). The rightmost column in Table 6.4 also demonstrates that the gap between 3L items and items of higher letters was also smaller in all cases for high Gf participants (-11%) compared to low Gf participants (-20%). We hypothesized that the gap to 3L items would be small when systematicity is on, but this was also assuming that everybody was taking advantage of systematicity. These results indicate that this is not the case – only high Gf participants are taking advantage of systematicity, and they only gain benefits from it at higher steps[19], because the systematicity takes time to emerge across steps in the item. This also relates back to a potential upper limit of binding capacity of four (Cowan, 2001; Halford et al., 1998). These post-hoc results demonstrate that true five-binding items may indeed be difficult, even for high Gf participants, and only through exploitation of strategies can they be made feasible. Combining the Swaps and ACT data indicates that binding capacity may be related to Gf with a two-fold explanation. The Swaps task indicates that high Gf participants are more likely to exploit systematicity; while the ACT task indicates that, when systematicity cannot be exploited, high Gf participants are also better able to cope with increased binding demands.

Thus, although the systematicity findings on the ACT and the Swaps task are in opposite directions, there does appear to be reasonable explanations for how the contrasting results occurred that is still harmonious with a relational integration perspective. In any case, once again, there was minimal evidence produced that attentional control relates to Gf (assuming that attentional control is properly operationalized by the increase in Steps). These explanations also demonstrate how important it is to carefully consider each task. The Swaps task was chosen because it seemed to exemplify the rearrangement (or ‘mental permutations’;

[19] Independent samples *t*-tests confirmed that there was a significantly higher systematicity advantage (i.e., the difference between *systematic* 4L+5L items and *non-systematic* 4L+5L items) for high Gf participants, as compared to low Gf participants ($t_{103} = 2.11$, $p = .037$) in 4-step items. The systematicity advantage for items of lower steps was not significant.

Stankov, 2000) aspect that was crucial to the success of the ACT. However, we have seen that in changing the task format to prioritise rearrangement of positions, the loss of the mapping aspect of the task has taken away the obvious fixed ordering that was required for low Gf participants to notice. A point that was discussed in Chapter II was Cowan's (2000) suggestion for measuring working memory: to truly measure and compare capacity, chunking opportunities should be clearly obvious to all participants *or* there should be no opportunity for chunking. This is an essential point because, as we have seen in this study, if the chunking opportunity is not completely obvious, only a subset of participants will take advantage of the chunking; and, as we have seen in the ACT study, this subset seems to have considerable overlap with the type of participants who tend to have higher capacity to deal with additional bindings. Without careful consideration of these chunking opportunities, apparently paradoxical findings can emerge, where both systematic and non-systematic presentations can relate to Gf. In all cases, high Gf participants are more capable of solving the problems, but without careful task analysis, the reason for their success may remain elusive.

It is also possible that the systematicity advantage (exclusive to high-Gf participants) discovered in these post-hoc analyses can be related to the induction aspect of Gf identified in Chapter II. This induction aspect is theorized to be somewhat independent from the relational integration aspect and the sole defining factor of Gf tasks, as compared to WM tasks. In the current study, we have observed how a WM task such as Swaps can have a potential inductive aspect, relating to the emerging systematicity. Thus, it is unlikely we can ever truly separate WM and Gf tasks even if we were certain of the overlap (relational integration) and the difference (induction), because they act upon one-another. Inductive processes 'exclusive' to Gf still supports relational integration through an increased sensitivity as to what must be bound in the representational system.

6.4.1. Conclusion

In the current chapter, we explored a task manipulation that aimed to separate attentional control from binding capacity. A theory throughout the thesis has been that relational integration (as measured through binding capacity, here) is related to Gf, where attentional control is not; but the two perspectives have been difficult to distinguish operationally. The Swaps task was an ideal measure for demarcating the two perspectives. Although there was some interaction between the two, the majority of variance in the Swaps task was associated with independent manipulations of capacity (through the number of letters involved in each problem) and attentional control (through the number of steps involved in each problem). However, as hypothesized, the variance associated with increasing demands on binding capacity were related to Gf, while the increasing demands associated with attentional control were not. A closer analysis on systematicity revealed this could at least partially be explained by high Gf participants taking advantage of systematicity, exploiting fixed elements to effectively reduce the capacity demands of higher-letter problems. On the surface, these results appear to go against the findings of the ACT (Chapter IV), but a comparative task analysis demonstrates that weaker participants may not have recognised the advantage of the fixed ordering in the Swaps, which was obvious and necessary for success in the ACT. This is a reasonable explanation considering that exploitation of systematicity in the Swaps task does involve an initial up-front cost associated with dissociating the fixed elements from the letter set. Thus, while the results are not perfectly in line with the hypotheses, the Swaps task nonetheless provided considerable insights above and beyond the previous studies. These insights are both theoretical, with the decoupling of attentional control from capacity; and pragmatic, demonstrating the importance of task analysis.

VII. STUDY 5: THE CHANGE DETECTION TASK

This chapter is published in Bateman, Ngiam, and Birney (2018). [Bateman, J. E., Ngiam, W. X. Q., & Birney, D. P. (2018). Relational encoding of objects in working memory: Change detection performance is better for violations in group relations. *PLoS ONE*, *13 (9), e0203848*]. There are changes to terminology and flow to fit the thesis.

The four studies explored thus far have used an experimental-differential approach to simultaneously answer the two core research questions: how can relational integration explain limitations in working memory and how do these limitations relate to fluid intelligence? One prominent assumption made by the binding approach is that bindings in active memory are inherently tied together by some relation connecting the bindings. If this is true, then it would suggest that memory for relations are strong but memory for details unrelated to the relations are weaker. This suggestion can be exploited by using a change detection paradigm where the task is for respondents to notice if a change occurs between two sets of similar stimuli (which may or may not actually be different). In most change detection paradigms (Rensink, 2002), the stimuli are temporally separated with a first, initial exposure to a “probe” display followed by a “test” display. The probe and the test displays may or may not actually be the same, and the participants’ task is to judge whether the displays are the ‘same’ or ‘different’. Participants often experience ‘change blindness’ during change detection tasks, even for seemingly large changes (Rensink, 2002). Of relevance to the current topic, if participants can more easily notice changes to relational aspects of a scene, but not to changes that maintain the relational aspects, it would indicate that active storage is indeed based fundamentally on relations.

Although change detection, and even the use of relations in change detection, has been explored in visual short-term memory research (Jiang, Olson, & Chun, 2000), the current study investigates this phenomenon from a working memory perspective in a large sample of undergraduate students. The large sample size allowed for particularly intricate analyses on learning effects, and other manipulations such as presentation time to be considered, to demonstrate that working memory is fundamentally based on relational integration. The introduction below outlines the change detection paradigm in further depth and explains how the current manipulations extends the literature.

7.1. Introduction to the Change Detection Task

Change detection paradigms have been employed to investigate visual short-term memory (Rensink, 2002). Using change detection, Jiang et al. (2000) discovered participants tended to remember individual objects in relation to the object's surroundings, even when the surroundings were task-irrelevant or when the target had been explicitly cued. Specifically, Jiang et al. had participants respond with either 'same' or 'different' to a probed target object which either did or did not change colour. The key experimental manipulation was whether the background to the object (consisting of additional objects) also changed or remained consistent. The backgrounds were irrelevant to the actual decision of the participant, but Jiang et al. found performance fell substantially if the background changed (e.g., if the background objects changed location or were no longer present on the test phase). This 'relational grouping' (encoding the configural relationships between objects into memory) has been shown to enhance recall (Rensink, 2000b; Ridgeway, 2006), suggesting that relational grouping is a necessary aspect of maintaining individual units of information (elements) in working memory (N. J. Cohen & Eichenbaum, 1993; Cowan, 2001; Halford et al., 1998; Oberauer, 2009a). The current thesis supports the suggestion that short-term memory is another description for the direct access region of WM and even these so-called 'visual short-term memory' tasks are demanding similar relational-based constraints. The current study aims to demonstrate this by employing the change detection paradigm to show that performance on change detection is fundamentally dictated by the level of relational grouping, and that the format of the task can enhance or detract from this tendency for relational grouping. Specifically, we examine the impact of relational grouping on change-detection performance by manipulating whether the change maintains the relational structure of the target group rather than changing the background stimuli (as in Jiang et al., 2000). Further manipulations to task format including the exposure times are considered, to determine how they influence the tendency to rely on relational information.

The short-term memory system is responsible for maintaining temporary information in a highly accessible state over a short period of time (typically in the realm of seconds and minutes), whereas *working memory* (Baddeley & Hitch, 1974) often refers to maintaining *and manipulating* information. This distinction is not often made in perceptual experiments (Luck & Vogel, 1997) where visual WM is the preferred term. Visual WM research involves brief exposure times (less than one second) (Vogel, Woodman, & Luck, 2006) and simple displays to assess immediate encoding performance while cognitive WM research typically allows participants to study elements (Cowan, 2001). As discussed in Chapter II, contemporary cognitive theories[20] of WM see the maintenance and manipulation of information inherently intertwined (Cowan, 2001; Oberauer, 2009a; Shipstead et al., 2016), such that capacity limits in WM are simply limits on how information is integrated into chunks of information (Cowan, 2001). Although WM is often defined by the *manipulation* of information, cognitive WM theories posit chunk-formation involves relational integration processes. Given the importance of this integration process to cognitive WM theories, the current study focuses on this chunk-formation.

[20] I use the term ‘cognitive theories of WM’ here to distinguish from visual short-term memory (VSTM) theories also related to the current study’s paradigm. Cognitive theories of WM include all the theories discussed in Chapter II, which theorize on the cognitive architecture of WM and consider how individual differences in capacity limits relate to higher-order processes. Conversely, VSTM theories are more concerned with explaining how visual information is encoded and capacity limits are almost exclusively related to the construction of the objects encoded (e.g., features, shape, orientation). In other words, the role of the individual is much more important to cognitive WM theories and rarely relevant to VSTM theories, but both domains of theories are interested in explaining how and why capacity is limited.

As described in Chapter II, Cowan (2001) suggests that chunks or the individual elements wherein are not directly related to the capacity of WM but are stored as a relation to some concept. For instance, recalling the sequence of letters “F-K-L” involves instantiating a relation to the concept of serial order. Similarly, Oberauer (2009a) proposes that information is maintained in WM by binding elements into a coordinated relational schema. For example, the recall elements F-K-L can be maintained through a schema of temporal order with F bound to temporal position x^1 such that: F^1, K^2, L^3. Halford et al. (1998) holds a similar ‘binding’ view of WM but puts an emphasis on the contribution of processing limits to the ability to instantiate new relations. For Halford et al., the capacity of WM is limited by the maximum number of elements that must be simultaneously considered to comprehend the relation that connects them. Although these theories (Cowan, 2001; Halford et al., 1998; Oberauer, 2009a) each have some unique aspects, they share the view that WM capacity is based on relational information rather than individual pieces of information. In these approaches, the ostensibly ‘un-manipulated’ maintenance of information is still subject to ‘processing-like’ limitations because the elements are stored via a common relation that must be instantiated.

Analogous perspectives in the visual short-term memory literature also highlight the importance of relational information. Vidal, Gauchou, Tallon-Baudry, and O'Regan (2005) suggest that relational information is gleaned from the visual display and form a ‘structural gist’. Changing a feature of a non-target (i.e., background information that is not part of the decision) changes the ‘structural gist’ and impairs change detection despite not being actually being operationally relevant to the target information (i.e., information critical to the decision). Similarly, Rensink (2002) proposes that relational information between a set of objects is pooled into a ‘nexus’ that contributes to higher-level decision-making (i.e., decisions about the group, rather than the object). The nexus is similar to the initial pooling of information in the structural gist process, though whereas the nexus exists as a separate source of information, the gist is bound with individual object information. The visual short-term memory approaches are quite similar, though the nexus (Rensink, 2002) suggests a more economical explanation, and accounts for the finding that it is easier to detect a change (among a group of non-changing objects) than it is to detect the absence of a change (among a group of changing objects) (Rensink, 2000a). Despite this difference, both the structural gist and nexus theories offer similar predictions on the importance of relational information, as do cognitive WM theories.

Considerable work has been devoted to determining the nature of storage and processing limits, and how relational information changes this capacity (Gabales & Birney, 2011; Miyake & Shah, 1999; Oberauer et al., 2003; Rensink, 2000b; Wilhelm et al., 2013). As noted however, the maintenance of elements through relations means that binding is an essential aspect to even basic storage-over-time tasks that have little higher-order processing. The impact of a relational integration theory of WM (based on binding capacity limits) on even simple storage-over-time experiments is understated, because WM theories are typically concerned with explaining the link between WM and higher-order abilities such as reasoning or problem-solving (Halford et al., 1998; Oberauer et al., 2008), rather than basic visual tasks. The current study was devised to strip things back to the basics of simply how information is maintained over time, supporting and extending on literature that suggests even this basic maintenance is fundamentally dependent on relational information.

7.1.1. Manipulations in the Change Detection Task

Jiang et al. (2000) demonstrated that change detection for a cued target worsened when unrelated background stimuli was altered, indicating that accurate memory for the identity of individual objects is influenced by the object’s seemingly irrelevant surroundings. One constraint on Oberauer’s (2009a) binding framework is that elements must be bound into a relation to be mentally represented in WM. A singular object can be bound in a unary relation of space but there is no frame of reference with which to compare changes. By also binding the target object's surroundings, altered relations between the object's surroundings cue the observer that a change has occurred. Indeed, Jiang, Chun, and Olson (2004) found that the poor performance associated with tampering with the surroundings could be attenuated by providing an invariant frame of reference (e.g., gridlines) as an additional context for the target to be bound alongside.

Although Oberauer's (2009) cognitive-relational WM can account for these results, both Jiang et al. (2000) and Jiang et al. (2004) involved brief exposures times (under 1 second) typically used when researching visual WM. Dent (2009) employed longer exposure times (2 seconds) in the realm of cognitive WM, manipulating whether changes to a target object were coordinate-only (a shift in position that maintained relations between objects) or categorical (a shift in position that violated the categorical relationship, e.g., above-of became below-of). Despite both types of changes being identical in veridical magnitude (in terms of change in visual angle), the categorical changes were detected at a higher rate than the coordinate changes. Dent's displays were simple in nature (only four objects per display) and changes were always a single object moving. In the current study, we similarly employed longer display times but investigated change detection with multiple *clusters* of objects, to better determine the effects of an integrated chunk of objects. If a cluster of objects is bound together, the rate of correct change detection should be drastically different depending on whether the objects move together or independently. Consider Figure 7.1. If we assume clustered objects are bound together, then we should see enhanced detection ability if the change occurs to a single object (the blue change in Figure 7.1A), because the test display has now arranged the cluster in a way that violates the bindings in WM. Alternatively, if individual objects are stored without bindings to other objects, the cluster change (red change in Figure 7.1B) should have enhanced detection, because it has three times as many objects violating the representation of the display in WM.

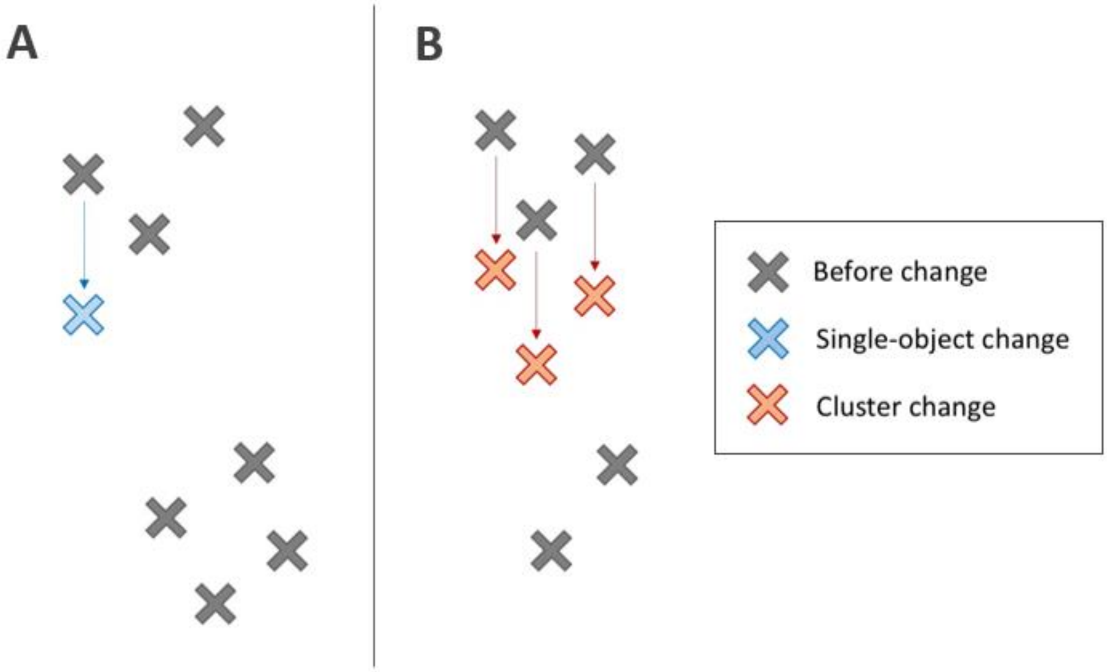


Figure 7.1. Example of relational encoding during the proposed change-detection task. Probe objects (grey crosses) are encoded as chunks of objects, due to proximity. The encoded relational information means a single-object change (indicated in blue, in panel A) would be easier to detect than a cluster of objects changing (indicated in red, in panel B), despite more objects overall moving in the cluster change.

In sum, if WM is primarily based on relations, we would expect a single-object change to produce a greater detection rate than the cluster-object change. Unlike Dent (2009), who focused on small set size displays and contrasted the position of two singular objects against one another, our displays involved large set sizes that clearly exceed the capacity of WM, but which were spatially arranged into clusters of objects. It was predicted that multi-object *cluster* changes, despite involving a change of a larger (surface) magnitude (i.e., more objects shift location providing more cues for detection), would be harder to detect than *single-object* changes as the spatial relation of the cluster is maintained. We designed the displays to encourage chunking of clusters: objects of the same cluster were the same shape (e.g., squares) and were closer in proximity to each other than to objects of other clusters (van

Lamsweerde, Beck, & Johnson, 2016; Woodman, Vecera, & Luck, 2003). This encouraged elaborated encoding: the high number of objects could be offset by grouping them into manageable chunks (Brady & Alvarez, 2014) that were clearly defined. This “encouraged chunking” allowed more control over participants’ approach to the problem, mitigating the use of unconventional strategies like chunking with the borders which would contribute to error outside the core manipulation (Cowan, 2001). Nonetheless, because participants tend to adapt strategies over time to suit the demands of a task (Lohman & Lakin, 2011), it was a real possibility that these unconventional strategies quickly become the preferred approach to dealing with cluster changes. That is, once participants recognize that half of the changes involve a cluster moving, they may shift their grouping strategy to not rely on bindings provided by an integrated cluster. To account for this, we consider the data at the item level, with the order of trials added as a predictor variable in our regression model. If working memory is fundamentally relational (Oberauer, 2009a), we would expect a large detriment for cluster change detection compared to single-object change detection, as cluster changes maintain the relation. However, over time, as participants learn to use adaptive strategies which help specifically with cluster changes (such as binding to the screen border), we expect the difference in performance between cluster and single-object changes to minimize.

Two additional manipulations were included to help explore the effects of relational grouping on the change detection task. These included two between-subjects variables: (i) exposure duration, and (ii) direction of single-object changes. These variables had the potential to damage the integrity of the core manipulation (cluster vs. single-object change detection), but the between-subjects variation meant we could easily cut certain conditions that did not function as expected. The rationale for these two variables are described below.

Exposure duration: Because Dent’s (2009) experiment was closest in nature to the current experiment, we also allowed participants multiple seconds to study the probe, as opposed to the sub-1s exposures featured in Jiang et al. (2000). Considering the increased set sizes of the objects relative to Dent (2009), we allowed an additional second (3s instead of Dent's 2s). Pilot testing indicated that participants still struggled to perform above chance at 3s exposure duration. However, it was of interest to determine if this chance performance was driven by certain manipulations (e.g., cluster changes being impossible to detect at brief exposures), so we varied probe durations at 3s and 5s, expecting that the increased study time in 5s would allow for more elaborate encoding strategies that accommodate cluster change detection.

Single-object change direction: Single-object changes can involve the target object shifting away or moving closer towards its cluster. Consider Figure 7.2. Changing whether 7.2A or 7.2B is the probe (and the other display is the test) varies whether the target is *distancing* or *uniting* relative to the cluster.

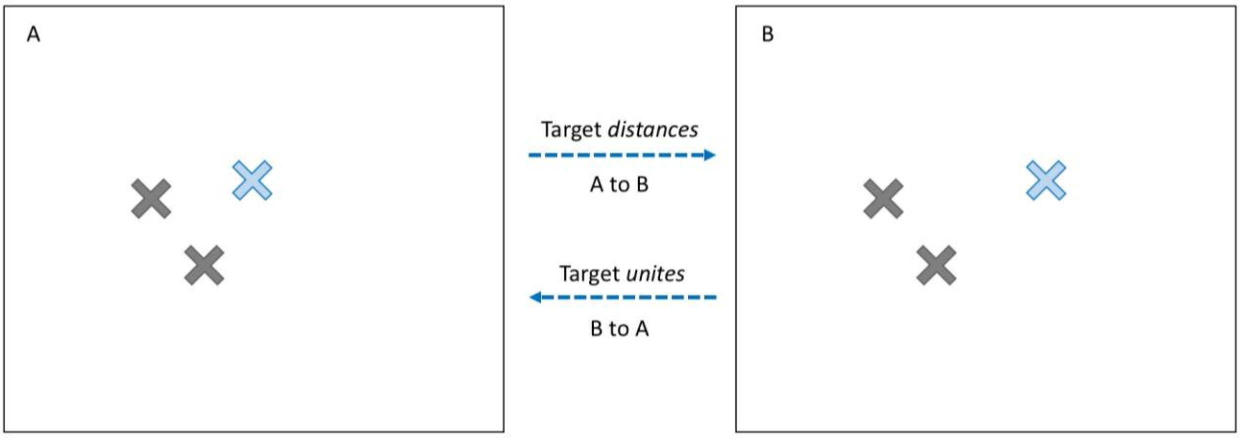


Figure 7.2. Example of the two types of single-object changes: *distancing* vs. *uniting*. If the display on the left (A) is the probe and the display on the right (B) is the test, then the target object has *distanced* itself from its cluster. Conversely, if 2B is the probe and 2A is the test, then the target object has *united* with its cluster.

According to cognitive WM theories (Cowan, 2001; Oberauer, 2009a), if all objects of a cluster are bound together, there is no particular reason to suspect that distancing or uniting should lead to different detection rates, because both changes violate the relation. However, if the cluster is initially easier to encode as a chunk (i.e., the probe is 7.2A), we would expect distancing changes to have higher detection rates than uniting, simply because the cluster binding was more often complete in the distancing condition. Because the objects are more dispersed in 2B, the spatial relation might not be as easily encoded (Brady & Tenenbaum, 2013) and as such, the violation of the relation is more likely to be missed because the relation was weakly encoded initially.

7.1.2. Hypotheses

The core goal of this experiment was to demonstrate the inherent reliance on relational information when performing a simple lower-order task like change detection.

H1: Changes which violate an encoded relational structure of a display will be easier to detect than changes which maintain the relational structure. Operationally, this means single-object changes will be easier to detect than cluster changes. This is the *target change* variable.

The remaining hypotheses were dedicated to the goal of determining how this inherent reliance on relational information manifests.

H2: The inherent reliance on relational information can be overcome with practice and experience. Operationally, the difference between cluster changes and single-object changes will minimize over the course of the practice items (i.e., an interaction between item order and target change).

H3: The inherent reliance on relational information will be more vulnerable when there is insufficient exposure time to perform more elaborate encoding. Operationally, the difference between cluster changes and single-object changes will be larger for 3s exposure duration compared to 5s (i.e., an interaction between exposure duration and target change). Although both types of target changes will suffer a detriment in detection performance in 3s relative to 5s, cluster changes should experience a larger decrease because there is insufficient time to incorporate non-intuitive encoding of clusters using other elements such as screen borders, while single-object changes can still be detected using the intuitive relational strategy.

H4: Uniting single-target changes will be harder to detect than distancing single-target changes, because the intended cluster is less likely to be encoded as a group due to the increased distance between each object.

7.2. Method

7.2.1. Participants

Undergraduate students participated in the study in tutorial groups as part of an assignment for their psychology course and were asked at the end of the study whether they consented to contribute their data to further research purposes. This method of recruitment was approved by the Human Resources Ethics Committee at the University of Sydney as an amendment to the ethics approval for the thesis project. Only the data of those who consented are presented here. In total, 952[21] first-year psychology students (70.2% female) at the University of Sydney participated. The mean age was 19.42 (SD = 3.22) years.

7.2.2. Measure and Procedure

Participants completed a change detection task, programmed using *Inquisit Lab 5* (Millisecond Software, 2017) and administered via desktop computer. Participants were tested in tutorial groups of 15-25. On each item, participants first viewed a *probe* image consisting of various shapes for either 3 or 5 seconds. Following a 3 second inter-stimulus interval, the *test* image was displayed, and participants responded whether this test image was the *same* (using the ‘A’ key) or *different* (using the ‘L’ key) to the probe presented previously.

[21] The large sample was the result of convenience. We acknowledge that this results in high power for the study, potentially exaggerating the results. As such, we reran the regression analyses five times using randomly selected subsets of the data (n = 200 each). Overall, none of the main effects changed significance during any of these subsets. Interactions occasionally fell out of significance, though this was more due to increased variability in the confidence intervals than the size of the effect itself (odds ratio).

Items were designed such that 10-12 objects were arranged on an invisible 10 x 10 grid, centred on the screen. Each space on the grid was 2 x 2 cm and each object was 1 x 1 cm. Objects could not fall in the outer cells of the grid but could appear on grid intersections. The objects were shapes of four kinds (circles, squares, triangles, crosses) and grouped into four clusters. The clusters were grouped both proximally and by kind of shape. That is, objects of the same group were closer in proximity to each other than to objects of other groups; and objects of the same cluster were all the same kind of shape (i.e., all squares). These design constraints were to facilitate grouping as a strategy to circumvent the otherwise large set size, allowing us to bias which groups were being formed by participants.

Participants first viewed task instructions which specified that the change would only ever concern location (objects moving, rather than changing identity) with demonstrations of both single-object and cluster changes. Twelve items were then administered as a ‘practice’ set (this set forms the analyses presented here). After the twelve practice items, a further 32 items were administered as ‘test’ items which introduced additional task manipulations and accompanying instructions. These additional manipulations turned out to be more impactful than anticipated and warped the results of the test set, making the data substantially more complex and challenging to interpret. For transparency, a brief summary of these additional manipulations can be found in Appendix C (along with descriptive results) but for now, the remainder of the chapter focuses on the practice set only, of which there were three core manipulations, described in turn.

Target change: Half of the items were *no-change* trials and the other half were *change* trials. The change always involved one or more objects shifting location by 1.5 spaces[22] (of the 10x10 grid; 2.5cm) in one of the eight cardinal or intercardinal directions.

[22] Pilot testing of different movement lengths indicated this was a sufficient degree of change to elicit responses above chance but below ceiling.

Half of the *change* trials (a quarter of all trials) were *cluster changes*, where a clustered group of objects changed location in the same direction together. The other half of *change* trials were *single-object changes*, where a single object changed location (i.e., a single object changed independently of its cluster, which remained the same). Figure 7.3 demonstrates the target change manipulation with the three types of trials (same, cluster change, single-object change).

Exposure duration: Participants were randomly allocated to a probe exposure duration of either 3 seconds or 5 seconds.

Direction of single-object changes: Participants were randomly allocated to a direction condition, such that single-object changes for half the sample involved the object *distancing* from its cluster while for the other half, the object *united* towards its group.

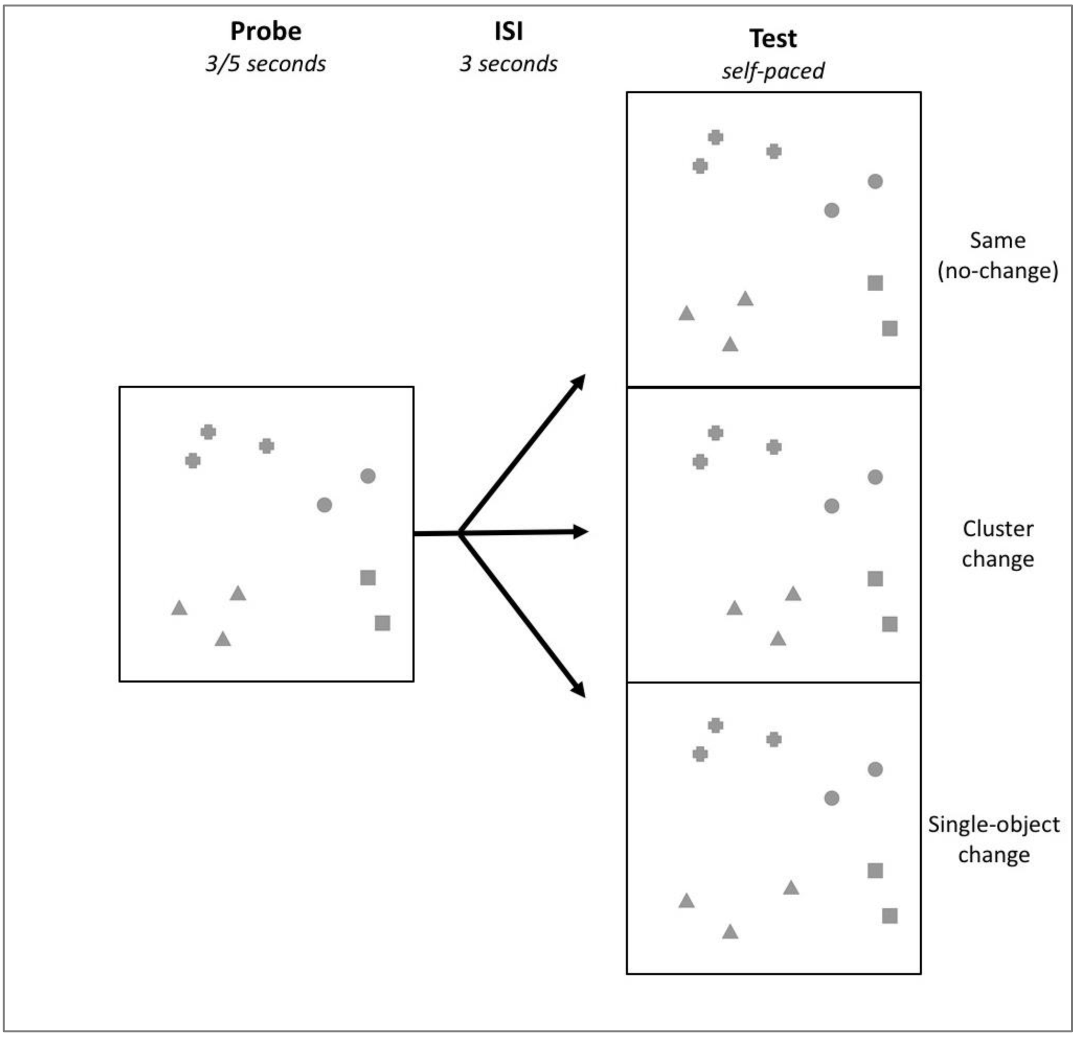


Figure 7.3. Example item demonstrating the three types of trials: same (no change), cluster change, and single-object change. In the example above, the target shape is triangle.

7.2.3. Approach to the analyses: Signal detection vs. proportion correct

Although it is common to use signal detection theory (Macmillan & Creelman, 1991) to form a dependent measure that accounts for sensitivity (such as *A'* or *d'*), because our core independent variable manipulated *type* of change, we would need to use the same false alarm rate (i.e., proportion of incorrect *same* trials) for both cluster and single-object conditions. This would mean every effect comparing cluster and single-object detection rates would be comparing two measures of which exactly half of each measure is perfectly overlapping with the other. This would raise multicollinearity and substantially reduce the size of any effect. Thus, instead of using a signal detection measure, we simply use raw proportion correct as the dependent measure. The main limitation of using raw accuracy, rather than a signal detection measure, is that signal detection considers the tendency for participants to favour one response over the other. For instance, if participants favour pressing ‘different’ over ‘same’, their raw accuracy for *change* trials will be high but at the cost of low raw accuracy for *no-change* trials. This would normally be a problem if we are only comparing accuracy of *change* trials to *no-change* trials, because this bias may not be apparent. However, because we are primarily comparing one type of *change* trials to another, it is less of a concern. In any case, there are also two further reasons to suggest that participants were not particularly biased. First, participants were explicitly told in the instructions that roughly half of the trials will contain changes and the other half will not. Second, the normal distribution indicates that the majority of participants were unbiased (six ‘same’ responses out of 12) or only slightly biased (seven ‘same’ responses). Figure 7.4 shows a histogram of how much participants favoured the ‘same’ response over the ‘different’ response. Although there is a lean towards ‘same’ responses (of, on average, 0.83 of an item), this is not surprising given the challenging nature of the cluster change trials. Thus, there is little reason to be overly concerned by using raw accuracy as a dependent measure.

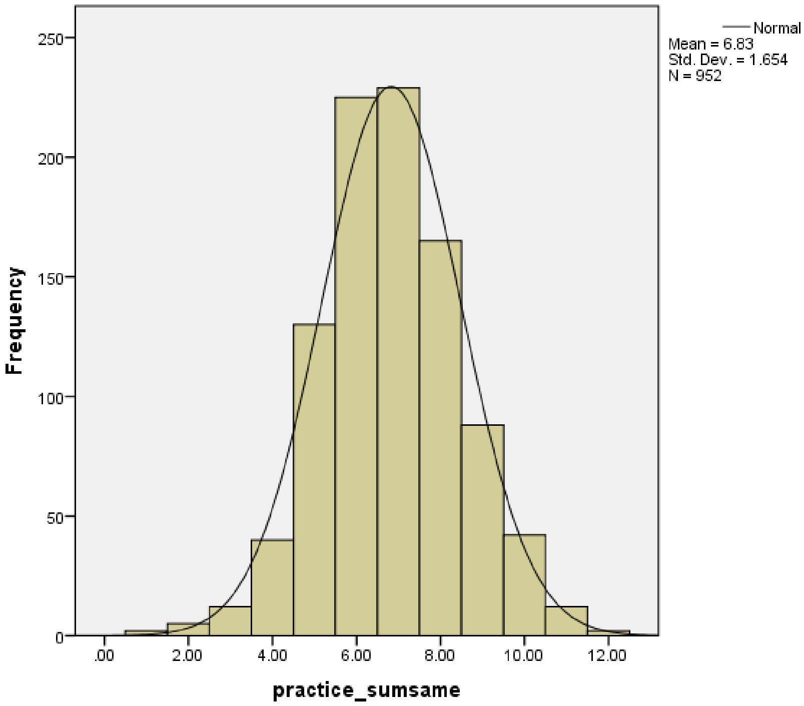


Figure 7.4. Histogram representing the number of ‘same’ responses each participant gave in their set of 12 trials, indicating a bias towards ‘same’ responses rather than ‘different’ responses (mode = 7). The bias is to be expected given the difficulty of detecting the cluster change items. The normal distribution indicates that participants were not overly biased, alleviating concerns over the use of composite accuracy scores rather than a signal detection score.

7.3. Results

7.3.1. Performance trajectories

Analyses of the practice set were conducted by modelling item responses using a mixed-effects logistic regression approach, to determine the influence of each variable alongside performance trajectories (trial order). The analyses were conducted using the ‘glmer’ procedure from ‘lme4’ (1.1.17) package (Bates et al., 2017) and performed with R version 3.5.0 (R Core Team, 2018). Plots were produced with the ‘sjPlot’ (Lüdecke, 2017) and ‘ggplot2’ (Wickham, 2009) packages. In total, 952 participants provided 11,436 data points for analysis (excluding same items: 5,718 data points). The overall proportion of correct trials was .819 for same items, .635 for cluster changes, and .729 for single-object changes. Figure 7.5 demonstrates these proportions split across direction and exposure times. Collapsing over direction and exposure time, the difference between change types was statistically significant (χ^2 = 365.64, *p* < .001), such that accuracy was greater in same trials than single-object trials (OR = 2.65, se = 0.14, CI95% [2.39, 2.94], *p* < .001), and (consistent with H1) single-object accuracy was greater than cluster accuracy (OR = 1.56, se = 0.09, CI95% [1.39, 1.75], *p* < .001).

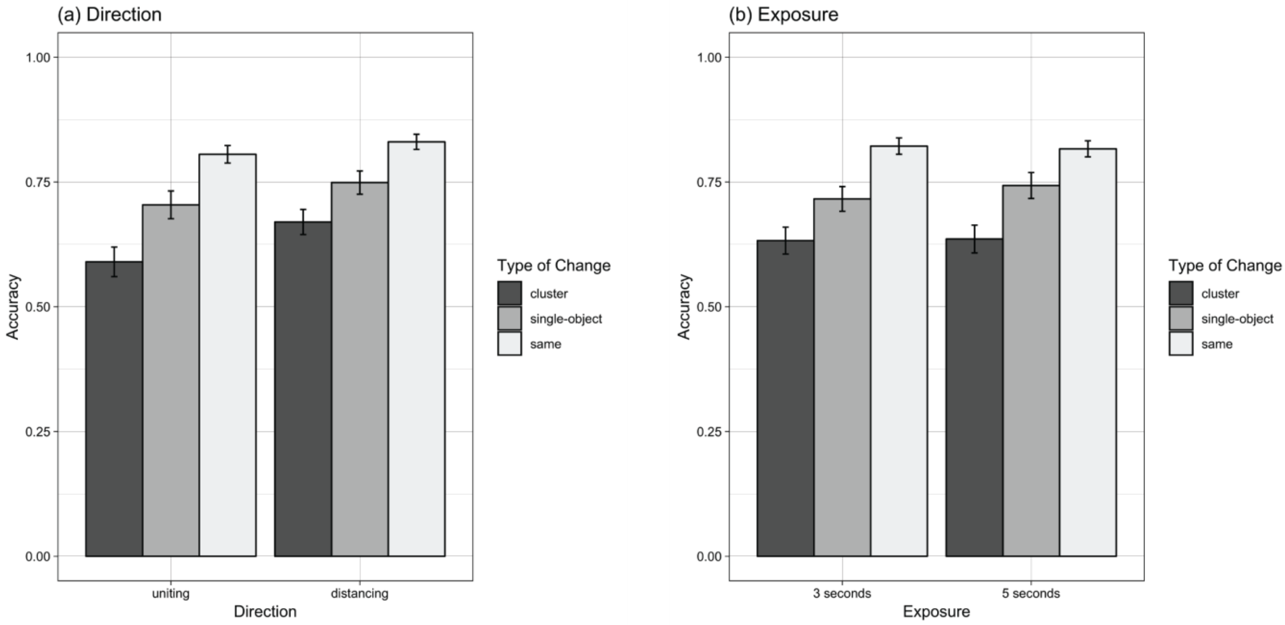


Figure 7.5. Average proportion correct for target change types in the Practice Set, broken down by (a) direction and (b) exposure. Error bars represent 2 x standard errors.

As the focus of our analyses is on differences between single-object changes and cluster changes, subsequent analyses of the practice set excluded *same* items. All variables (i.e., target *change* type (single-object, cluster), single-object *direction* (distancing, uniting), *exposure* duration (3s, 5s), and *trial order*) and their interactions were regressed on accuracy. The regression coefficients are reported in Tables 7.1 and 7.2 and trial-order trajectories are demonstrated in Figure 7.6.

There was no main-effect for *exposure* (OR = 1.13, se = 0.08, CI95% [0.98, 1.30], p = .101) and *exposure* did not interact with any of the other variables including *change* type (contrary to H3). There was a significant main-effect for *direction* (OR = 0.77, se = 0.06, CI95% [0.67, 0.89], p < .001), such that accuracy was higher for distancing items than uniting items. Although the *direction x change* interaction was not significant (OR = 1.17, se

= 0.14, CI95% [0.92, 1.48], p = .200); the three-way *direction x change x trial order* interaction was (OR = 1.08, se = 0.04, CI95% [1.01, 1.16], p = .026). As can be seen in Figure 7.6a, this pattern of results indicates that single-object performance is worse for uniting than distancing (consistent with H4) but only for earlier trials: by the end of the 12 items, the difference between distancing and uniting for single-changes has closed.

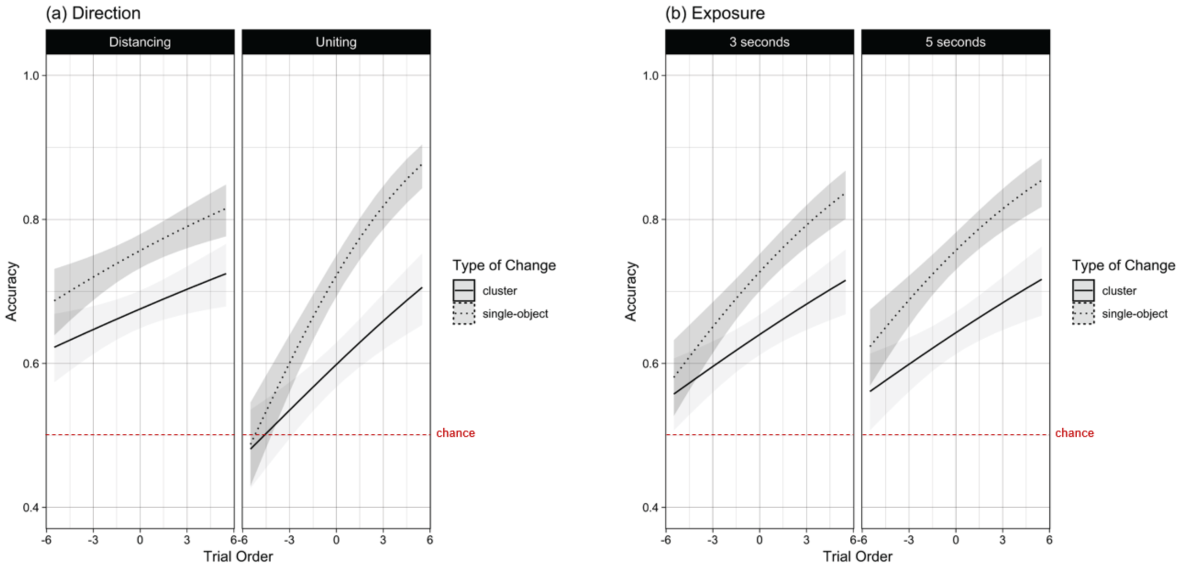


Figure 7.6. Model plots of interactions (conditional on all other variables) for (a) Direction, and (b) Exposure. Shaded areas 95% CI.

Trial order was a significant predictor of accuracy (OR = 1.07, se = 0.01, CI95% [1.39, 1.75], *p* < .001), with participants becoming more accurate in detecting change across the 12 items. Overall, the trajectory was more pronounced for single changes than cluster changes (OR = 1.06, se = 0.02, CI95% [1.02, 1.10], p = .001). Although the presence of an interaction is line with H2, Figure 7.6 demonstrates that the interaction was in the opposite direction to the one predicted. That is, the interaction was primarily a result of detection rates improving faster for single changes as opposed to cluster changes closing the gap to single changes (as was hypothesized). Thus, contrary to H2, the two types of changes had more similar performance to begin with and grew more dissimilar over the course of the 12 items. As seen in Figure 7.6a, this interaction effect was more pronounced for *direction* being uniting rather than distancing (OR = 1.08, se = 0.04, CI95% [1.01, 1.16], p = .026). Simple-interaction analyses (on this three-way interaction) indicated that the single change trajectory was significantly more pronounced than the cluster change trajectory for uniting items (OR = 1.10, se = 0.03, CI95% [1.05, 1.16], p = <.001) but not for distancing items (OR = 1.02, se = 0.02, CI95% [0.97, 1.07], p = .460); but again, as seen in Figure 7.6a, neither type of direction resulted in the hypothesized effect (H2: single and cluster performance becoming closer over the course of the 12 items).

Table 7.1. *Regression coefficients for all main effects and interactions, averaged over direction.*

	combined			
	Odds Ratio	CI	std. Error	p
Fixed Parts				
(intercept)	**2.24**	**2.09, 2.40**	**0.08**	**< .001**
direction (uniting vs. distancing)	**0.77**	**0.67, 0.89**	**0.06**	**< .001**
exposure (3s vs. 5s)	1.13	0.98, 1.30	0.08	.101
change type (single vs. cluster)	**1.62**	**1.44, 1.82**	**0.10**	**< .001**
trial order (mean centered)	**1.10**	**1.08, 1.12**	**0.01**	**< .001**
direction x exposure	1.18	0.89, 1.56	0.17	.255
direction x change	1.17	0.92, 1.48	0.14	.200
exposure x change	1.16	0.92, 1.47	0.14	.208
direction x trial order	**1.09**	**1.05, 1.13**	**0.02**	**< .001**
exposure x trial order	0.99	0.96, 1.03	0.02	.671
change x trial order	**1.06**	**1.02, 1.10**	**0.02**	**.001**
direction x exposure x change	1.25	0.78, 2.00	0.30	.357
direction x exposure x trial order	1.02	0.95, 1.10	0.04	.540
direction x change x trial order	**1.08**	**1.01, 1.16**	**0.04**	**.026**
exposure x change x trial order	0.99	0.92, 1.06	0.04	.823
direction x exposure x change x trial order	1.09	0.94, 1.25	0.08	.256
Random Parts				
τ2, subject	0.371			
N, subject	952			
ICC, subject	0.101			
Observations	5718			
Deviance	6221.116			

Note: All variables are mean-centred.

Table 7.2. *Regression coefficients for all main effects and interactions, split by direction.*

	Uniting (reversed)				Distancing (standard)			
	Odds Ratio	CI	std. Error	p	Odds Ratio	CI	std. Error	p
Fixed Parts								
(intercept)	**1.96**	**1.77, 2.18**	**0.10**	**< .001**	**2.55**	**2.32, 2.81**	**0.12**	**< .001**
exposure (3s vs. 5s)	1.23	1.00, 1.51	0.13	.053	1.04	0.86, 1.26	0.10	.711
change (single vs. cluster)	**1.75**	**1.47, 2.08**	**0.15**	**< .001**	**1.50**	**1.28, 1.76**	**0.12**	**< .001**
trial order (mean centered)	**1.14**	**1.12, 1.18**	**0.02**	**< .001**	**1.05**	**1.03, 1.08**	**0.01**	**< .001**
exposure x change	1.30	0.92, 1.84	0.23	.138	1.04	0.76, 1.43	0.02	.802
exposure x trial order	1.00	0.95, 1.06	0.03	.888	0.98	0.94, 1.03	0.02	.436
change x trial order	**1.10**	**1.05, 1.16**	**0.03**	**< .001**	1.02	0.97, 1.07	0.02	.461
exposure x change x tr. order	1.03	0.93, 1.15	0.06	.540	0.95	0.87, 1.05	0.05	.309
Random Parts								
τ2, subject	0.363				0.377			
N, subject	426				527			
ICC, subject	0.099				0.103			
Observations	2556				3162			
Deviance	2843.202				3377.964			

Note: All variables are mean-centred.

7.4. Discussion

The current study was devised to reinforce the basic suggestion that WM is inherently relational. Additionally, we tested task parameters (probe exposure duration, direction of change) and considered them alongside an analysis of trial order to assess how the basic relational change detection finding (that changes that maintain relations are harder to detect than changes that violate relations) manifested over the course of the task. The core manipulation was successful: single-object changes which violate the relation of grouped objects were more likely to be detected than cluster changes where the relation is maintained. The study thus contributes further evidence to the current thesis that WM stores and maintains information through relations. These results are consistent with Dent’s (2009) findings on singular objects that categorical changes (changing relation) are easier to detect than coordinate changes (maintaining relation). Thus, although grouping efficiently maximises the amount of information that can be stored at any one time (Cowan, 2001), the present data indicates that this can come at a cost: visual changes where the relation between the changed objects is maintained can be missed. This effect is demonstrated despite cluster changes involving changes of a larger veridical magnitude (overall, more objects change location) than single changes.

Even though change detection tasks are typically seen as visual STM paradigms, the cognitive-relational WM built towards in this thesis (Cowan, 2001; Oberauer, 2009) applies suitably to these results also: participants encoded objects as part of a larger structure, rather than situating individual items at particular coordinates. This is also consistent with Jiang et al.'s (2004) finding that changing task-irrelevant objects hinders performance not due to a general interference effect but because this disrupts the structure of the display.

A large sample that was unfamiliar with the task allowed us to assess naïve approaches and how that changed across trials. Cluster changes were initially difficult to detect but participant performance improved rapidly over the 12 trials. Although uniting cluster changes were initially more difficult than distancing cluster changes, both uniting and distancing cluster changes had similar levels of performance by the end of the task. This suggests that participants were becoming aware of the types of changes to expect (both in terms of cluster changes and in their respective direction), and potentially changing their approach to the task as a result. Although performance for cluster changes was initially worse, single-object change detection performance improved at a faster pace. It must be cautioned that this interaction was qualified by a three-way interaction with the change type and trial order suggesting this improvement was faster for uniting trials than for distancing trials (Figure 7.6a). It appears that uniting single-object changes are initially more difficult to naïve participants than distancing single-object changes, indicating there may be effects of ease of initial encoding (Brady & Tenenbaum, 2013). When objects are more disperse, it may be more difficult to form an accurate relation compared to when the objects form a tighter structure. The violations with the uniting changes are then less likely to be noticed, not because they act as fundamentally different relations, but because they have never been encoded as the same chunk in the first place. Interestingly, participants improved in performance on *uniting* single-object changes faster than those detecting *distancing* single-object changes, indicating that these encoding effects are quickly overcome with task familiarity. That is, once participants were aware of the nature of the groups (all the same shape), they were able to encode the target as part of the cluster despite the increased distance. This theory could be confirmed by employing another condition where the target object starts out at distance from the group (like the uniting condition) but moves even further away. Thus, this condition would still be 'distancing' but the initial chunk to encode would also be distant. Alternatively, or in addition, the improvements may be a result of increasing understanding of task requirements and consolidation of more effective strategies, which may be prone to related individual differences that have not been explicitly investigated here.

We found no difference in overall change detection performance or learning trajectory when comparing display exposure times (3 vs. 5 seconds). This suggests that 3 seconds was sufficient time to consciously encode chunks despite having as many as three times the number of objects as Dent's (2009) displays. This is consistent with Rensink's (2000b) finding that 12 items can be processed in approximately 1.5 seconds. The grouping cues of proximity and shape identity likely aided pooling of the objects (Rensink, 2002). If this is the case, then it is unlikely the extended probe duration equated to elaboration. Because both single-object and cluster performance improved over the course of the test, it is also possible that two levels of structure were formed simultaneously: one level encoding relations between items and one level encoding relations between clusters; with 3 seconds being sufficient to encode both levels. Hence, both single-object and cluster performance improved over time (albeit with single-object performance improving faster), because both levels of structure were being fine-tuned over the course of the task. If a lower exposure duration (e.g., 1 second) produced different trajectories (e.g., single-object performance improves, but cluster performance does not), it would confirm that two levels of structure are present.

It should be cautioned that the learning trajectories presented here are based on naïve participants and limited to 12 trials. Because performance did not reach asymptote in either single-object or cluster conditions, it is possible that performance trajectories could have continued or changed. Nonetheless, what we have demonstrated is that the learning trajectories of the two types of changes start out similar and quickly grow dissimilar (standard errors between the conditions generally lose their overlap by the third trial). Although we cannot say if this trend continues past 12 trials, it does indicate that initial learning of the task widens the gap between detection rates of single-object and cluster changes. Although this goes against our initial hypothesis (that participants would start with a performance advantage for single changes over cluster changes), it does instead support a conclusion that the detection of changes which violate relational structures is *learned* more intuitively than changes which maintain relational structure.

7.4.1. Conclusion

It is clear that structure and relation are critical to memory (Cowan, 2001; Halford et al., 1998; Oberauer, 2009a) and form the cornerstone of higher-order intelligence (Gentner, 2003; Krumm et al., 2009; Oberauer et al., 2008). Oberauer (2009) suggests that representations are only maintained within immediately accessible memory by the binding of an individual element to a context within a relational schema. As a result, actively maintaining elements is dependent on relational information. Single-object change detection was better than cluster change detection, but we cannot conclude that memory is entirely limited by this dependency, as performance on single-object changes was not close to ceiling and performance on cluster changes was above chance from the very beginning of the task. Although it is likely that relational information was still being used even during this naïve cluster change detection in some way (e.g., borders or nested clusters), we cannot specifically say we have evidence for a complete dependency on relational information in WM. Because our experiment was based on changes in spatial position, we also cannot necessarily generalize these findings to other visual properties or verbal information. Alterations to the task procedure (such as informing participants of the target shape) or measurement methods (such as incorporating biometric measures like eye tracking) may prove more conclusive. Nonetheless, the present results, together with a cognitive approach to WM (extending on more perceptual accounts of visual WM), produces interesting implications for our understanding of the process and constraints involved in grouping spatial information. The current results indicate that grouping information is an effective way to bypass capacity limits, but it comes at a cost: changes that maintain the relational structure of the display are more likely to go undetected. Multi-object groups can shift unbeknownst to participants if their spatial relation is maintained. It appears that maintaining information in WM is dependent on the relations that connects that information, as a single object changing independently of its relation is conspicuous.

VIII. GENERAL DISCUSSION

8.1. Discussion

Ultimately, the goal of this project was to demonstrate that working memory (WM) can be best understood as a system for relational integration, the process by which independent representations in memory are connected via their place within a common relation (Oberauer, 2009; Halford et al., 1998). Traditional storage-based theories of WM (e.g., Baddeley & Hitch, 1974) require significant amendments to explain the pervasive ability to chunk (Cowan, 2001). A relational integration theory respects the importance of chunking by building its foundation around the connections that bind elements together. WM enables representations to be stored over time but in itself is not a system based on the capacity to store representations, but to connect them. In this way, WM capacity is not restrictive, but permissive.

Relational integration theories align both with theories of higher-order cognition that explain abstract reasoning performance (Hummel & Holyoak, 2003) and beyond, to theories of comparative cognition that asks what make humans unique (Gentner, 2003). Despite the overlap with theories outside WM and despite the intuitive explanation of chunking, distinguishing the relational integration view as a pre-eminent theory of WM has been challenging (Cowan, 2017). This is in no part due to the difficulty in demarcating it from other general views of WM where task performance is restricted by broadly construed attentional processes. To aim in this distinction of relational integration, the current thesis looked to answer two fundamental questions (Conway et al., 2007) in WM theory: (i) what limits capacity in working memory, and (ii) how these limitations are related to performance on fluid intelligence (Gf) tasks. To accomplish this, we employed an experimental-differential approach (Birney & Bowman, 2009; Deary, 2001), where tasks are experimentally manipulated to vary in their theorized cognitive demands, then compared to established measures of Gf to determine which manipulations best explain the well-document overlap between WM and Gf. Overall, performance on our experimental tasks was consistently and concomitantly influenced by the changing demands on relational integration, answering the first research question on capacity limitations. Direct access capacity, as measured through the number of bindings that had to be simultaneously established, frequently linked to accuracy, with more bindings requiring higher demands and reduced performance. However, we also frequently found performance costs involved in varying attentional control demands, at times to a similar extent as the binding capacity demands. Critically however, the answer to the second research question, relating task manipulations to Gf, was regularly in favour of a relational integration view, with varying demands on binding capacity specifically and concomitantly linked to Gf task performance. The core of the remaining Discussion reflects on the implications of these findings.

8.1.1. Measuring Working Memory through Relational Integration: Capturing true binding capacity

A frequent challenge for WM researchers has been the measurement of storage capacity in a way that accounts for the ability to chunk information. The complex span paradigm attempts to solve this by using intermittent unrelated processing that hopes to pull attention away from conscious chunking efforts. The current thesis indicates that this unrelated processing is unlikely to fully prevent chunking but rather, simply taxes an additional attentional control demand that is (in some ways) related to capacity limits. We frequently found evidence for ‘true’ binding capacity limits of around three to five bindings. As examples, these capacity limits emerged in the Latin Square Task, with challenging 4D items; Swaps performance, with three to five bindings capturing a wide range of performance; and the Relation Monitoring Task, where relations usually involved three bindings. In all cases, these tests of binding capacity were sufficient to both extract systematic variance between individuals in the sample and link performance to higher-order Gf. The frequent, unrelated processing in the complex span paradigms is an indirect way of capturing binding demands, because chunking is not prevented by the nature of the elements, but (one hopes) by the format of the task. This means the complex span is unreliable in capturing WM because it allows for variation in the chunking strategy, which Cowan (2001) recommends should be limited to maximise variation attributable to the number of active bindings – a higher fidelity measure of actual binding capacity. This thesis has frequently argued that inductive components of a task (including development of strategic knowledge and on-task learning) more closely align to Gf than WM, and may represent the unique component of Gf that is independent from WM. Thus, although induction allows for frequently observed overlap between WM and Gf tasks, it obfuscates the interpretation of capacity limits in WM.

We were more often successful with manipulations of tasks that required every element to be simultaneously involved in the process or the outcome of the task in such a way that the elements could not be chunked. This premise was put forth generally by Cowan (2001) and shared by the more specific Relational Complexity (RC) theory (Halford et al., 1998). Often, we employed RC theory in the task analyses to ensure that every element had to be independently represented in the solution process. For some tasks, this worked well but it was not without limitations. In the RMT and LST, the problem solution is the outcome of a relational instance requiring two to four independent bindings (in line with RC theory). RC theory overall predicted the performance results for these two tasks quite well but rarely told the full story. Eye tracking analyses in the LST indicated that distractor cells (unrelated to the RC of an item) factored into item success, while the novel addition of the *ascending* condition in the RMT proved to be substantially more challenging than *same* despite having similar theoretical RC demands. These exceptions do not necessarily detract from the RC theory, which clearly cautions that RC is only valid when all other demands are controlled. However, these results do indicate that there are frequently times when increases in RC level can produce unintended interactions that make it impossible to truly control for all these demands. In the LST, the theoretical RC level increases linearly and in even intervals. In practice, the change required in going from 3D to 4D is substantial and catalytic to performance, while the change going from 2D to 3D is largely negligible, at least in a university population. In this case, this difference was theorized (in Chapter III) to be a result of interaction with a shift in strategy required for 4D items (from a shape-based to a dimension-based strategy). Therefore, despite the LST being a task borne from RC theory, RC theory failed to explain a majority of the story.

Different troubles for RC occurred in the Swaps task, which was highly demanding for participants, despite the core task involving fewer elements (typically three) than most set sizes of the complex span (which range from two to seven). This demand was actualized in the Swaps task via the constant rearrangement of letters, frequently requiring letters to be bound separately and simultaneously. Although each individual swap only ever involved two letters, it was argued that the format of the task necessitated multiple independent bindings because there was rarely a chance to take advantage of fixed bindings beyond a single step (only in the systematic condition). This (letters) variable ended up being remarkably successful in performance demands, yet RC theory could not be applied because the outcome of the task was not a clear outcome of a relational instance requiring independent bindings. Rather, the task itself was loading on binding processes through frequent binding and unbinding and the outcome was conceptualized in terms of the number of bindings in the direct access region (Oberauer et al., 2007; Oberauer, 2009a). Thus, although Cowan’s (2001) theory of chunking capacity is far more general than RC theory, it may be more applicable to a lot of the studies observed throughout this project, where demands were frequently placed on the binding process itself, rather than the outcomes of the binding process. Overall, we still preferred to frequently employ the RC theory because of its commendable specificity on complexity and instantiation, but this also led to its insufficiency when it could not comprehensively explain results in a way that a more general model could. Together though, this indicates that specifying factors involved in the binding process (e.g., the strength or flexibility of the bindings) could lead to meaningful task analyses that allow us to re-evaluate binding capacity demands throughout a task.

A clear recommendation that comes from this research is to follow Cowan's (2001) advice: minimize variation in chunking and maximize variation in the chunks. While this has always been the case, the current research presents several successful examples of this principle. The RC metric (Halford et al., 1998) ended up being most useful for identifying situations where bindings could be systematically chunked down. The subtle difference between access-fixed and access-random in the ACT was an example of systematicity. It required no additional instructions, yet precisely captured a component shared by WM and Gf, allowing us to conclude that only those high in Gf could deal with the additional bindings required by access-random. However, the results were not always this clear. In the Swaps task, which focused solely on the rearrangement of bindings, systematic items produced the strongest link to Gf, but that appeared to be because high Gf participants were taking advantage of the systematicity resulting in variance related to induction, rather than (only) binding. In this case, there was a failure to minimize the variation in chunking (strategy), which resulted in inflation of the WM-Gf correlation for the systematic condition: the induction component theorized to be unique to Gf was bleeding into the WM task. Although, at a task-level, a higher correlation may seem ideal; at a condition-level, it makes it harder to conclude the processes underlying WM. Thus, it is important to be vigilant with task analyses and consider variation both in chunking (strategy), and in the chunks themselves (Cowan, 2001).

8.1.2. Demarcating attentional control demands from binding capacity in the prediction of Gf

One of the difficulties in measuring binding capacity is that the active attention placed on the direct access region means predictions often coincide with attentional control theories. The studies in this thesis regularly tried to distinguish the two theories (despite the acknowledgement that they do overlap). In Chapter III on the LST, ‘dynamic completion’ was theorized to reduce attentional control demands by allowing participants to offload partial solutions. This condition successfully maintained the relationship to Gf, theorized to be because it still captures binding capacity involved in integrating the elements, despite the reduction in variance associated with attentional control demands.

In Chapter V, the Relation Monitoring Task (theorized to be a pure measure of relational integration) was validated by considering binding and attentional demands separately. While binding demands were captured through varying performance via the match type, the attentional control demands rarely led to changes in performance. One of the unknowns surrounding the task was whether strings needed to persist between arrays, in order to ‘reduce the amount of new information’ to be appraised (Oberauer et al., 2008). Although theoretically this seems reasonable (to load more highly on relational integration, rather than scanning), there was yet to be any experimental validation that this was necessary. If it turned out that the task performed worse as a predictor of Gf without string-preservation, it would imply that the task’s contribution to Gf is less about relational integration and more about attentional demands, such as scanning. Our results were encouraging for a relational integration perspective, with both versions of the task (string-replace and string-preservation) showing strong predictions of Gf over-and-above typical WM tasks (complex-span and *n*-back). The string-replace version did predict a slightly higher amount, attributable to a unique attentional component. Thus, although these results are still consistent with theories of attentional demands, they nonetheless appear to be non-essential demands in predicting Gf. We found similar results when employing interference (duplicated digits across the arrays): non-essential but a slight, significant unique contribution.

Finally, Chapter VI on the Swaps task was designed from the start with the aim to directly distinguish attentional control from binding capacity, via the independent manipulations of steps and letters. We found both steps and letters contributed to performance, but only increases in letters was related to increases in the correlation to Gf.

The most fitting conclusion to these findings then, is that both relational integration and attentional control can uniquely explain task performance. Each appear to produce independent contributions to task demands. However, only relational integration uniquely overlaps with Gf. The times when attentional control overlaps with Gf are either marginal or non-essential contributions (as in the RMT string-replacement and interference), can be explained through an overlap with binding capacity (as in the ACT), or can be explained through inductive processes related to on-task learning (as in the complex-span). Despite all this, it must be cautioned that the scope of this thesis clearly aligns with a relational integration view and thus, has inevitably been somewhat insufficient in representing attentional control to the same degree as relational integration. With that said, relational integration has itself been underrepresented compared to attentional control in the wider literature[23], so the hope for the studies presented here is that they provide a compelling reason to consider relational perspectives of WM in more depth.

[23] A simple comparison of search terms demonstrates this. In PsycINFO, as of December 2019, “relational integration + working memory” has only 27 results, while “attentional control + working memory” has 479. At least some of this disparity can be attributable to the (current) lack of established terminology for relational integration and the tendency to not use it in the title or keywords of articles (which is in itself a problem for developing the theory).

8.1.3. Measurement issues and additional contributions

Although the primary purpose of this thesis was to contribute to the two core research questions (which were answered in the prior two sections), there are a number of other points worth remarking on, as they provide additional contributions that may prove insightful to researchers yet did not have a place within any individual chapter. This includes unpublished data from additional studies that did not warrant a place within the main chapters of the thesis and a discussion to address pressing measurement issues that are universal to the entire thesis (and the WM literature as a whole).

In all, one of the most powerful findings was simply the raw correlation between the RMT and our Gf factor at r = .61. Considering the complexity of Gf tasks like the APM, which have still not been conclusively deciphered, it is remarkable that such a comparatively simple task like the RMT can predict nearly 37% of the variance in APM. Although the RMT and the APM take about a similar time to administer (20 minutes), a full WM and Gf battery can take upwards of a full day of testing. Further, the RMT is, on the surface, a simple matching task. Some evidence suggests that complex Gf tasks induce more test anxiety than ostensibly simpler tasks (Gimmig, Huguet, Caverni, & Cury, 2006). This is particularly relevant considering that ‘matrix-style’ tasks (like Raven’s) are well known intelligence tasks used in recruitment and aptitude testing (Carpenter et al., 1990). Thus, perhaps one of the most practical implications for the project has been to validate and promote the use of the RMT as an assessment both within and beyond the WM research body.

The next point relates to measurement issues. Across the studies employed, a wide variety of tasks were employed. Some had conditions that bordered on too easy (LST) and some too hard (Change Detection), while others were just right (RMT, Swaps) with a range that seemed to capture the extent of abilities in our university-level populations. In all cases, we operationalized WM demands in these tasks with accuracy. As seen in the LST-DC, even conditions with very high performance (near or at ceiling) could still be insightful if what variance remained was meaningful (i.e., the items demanding the most relational integration). The RMT and Swaps task had more ideal difficulty ranges, and yet the conclusions of how the variance within these tasks linked to the variance within Gf tasks were not necessarily easier to interpret than in the LST. Better psychometric properties does make for more powerful (statistical) effects (as described earlier with the RMT), but careful task analyses are still required to interpret these effects. In all, what this indicates is that a task with clearly defined cognitive demands will be more meaningful than an ambiguous task, even if the ambiguous task has more ideal psychometric properties. Of course, a bare minimum of variance is required to provide any meaningful individual differences (i.e., perfect performance across all participants is not useful except to prove that the population is beyond the threshold of task demands). And of course, the ideal is to both clearly define tasks and to achieve good psychometric properties, but even a psychometrically ideal task is not useful without a meaningful decomposition of the cognitive processes involved.

Another extreme example of the need for careful measurement of accuracy comes from additional data from the Change Detection task. As mentioned in Section 7.2.2, the data presented in Chapter VII was from the ‘practice’ set of the task (12 items), and there was an additional ‘test’ set of data with additional items (32 items). The test data did not make it into the main analysis, because it included an overly powerful experimental manipulation that pervaded the other analyses. This manipulation was an ‘ignore instruction’ that appeared either before or after exposure to the probe display and instructed participants to ‘ignore’ any changes that occurred to a certain type of shape (e.g., “triangle”). The intention with this manipulation was to determine the relative difficulty of binding vs. unbinding elements in the direct access region. The theory was that pre-probe instructions would result in higher performance than post-probe instructions, because there is additional difficulty in selectively unbinding one cluster (post-probe) as opposed to never binding the cluster in the first place (pre-probe). Unfortunately, this instruction timing variable was also confounded simply be the fact that pre-probe instructions reduced the set size to be remembered (from 10-12 to 8-9). Thus, although the hypothesis was ‘correct’ (pre-probe had higher performance than post-probe), this could have simply been a set size effect. More problematically was that this manipulation turned out much more powerful than anticipated and warped the outcomes of the other manipulations (including the core target change manipulation). This variable could not be averaged over (as was done in other aggregate manipulations like in the LST) because this variable contributed such extreme variance to performance that it rendered most other effects non-significant. That is, the timing of the instruction was so powerful that it had to be included in every analysis just to explain the large variance that would otherwise have been attributed to random error variance. This also reduced the item size representing each effect to unreasonably low amounts (k = 2, in the worst case) meaning some hypotheses simply could not be resolved. In this case, the lack of a careful task analysis that could have identified the demands in the timing conditions crippled the test data. For full transparency, the analyses conducted on the test data are provided in Appendix C. Although this manipulation in the Change Detection task suffered the most, aggregation caused some issues elsewhere. In the Swaps task, the systematicity variable was difficult to decipher (on aggregate level) because it interacted both with the Letters and the Steps manipulations (i.e., the effect of systematicity only really emerged at 4-step items, and only in 4- and 5-letter items). In the case of the Swaps task, this was less damaging to our interpretation because the other manipulations (Letters and Steps) were strong enough to allow for further task breakdowns to provide insight (as was done, i.e., in Section 6.4 deconstructing the systematicity advantage in each condition). A solution to these issues is to not overextend the experiment to include so many manipulations. In the case of the Swaps task, the systematicity variable was a necessary inclusion as an outcome of the task analysis conducted on the Letters variable. In this case, systematicity was going to influence the results regardless of whether we included it or not, so we opted to have full control over it by manipulating it as part of the experiment script. The reasoning for the Change Detection task was more fallible, as the manipulations were more a result of being tempted by the large sample size (*n* = 900+) than via a careful task analysis. In any case, a better (potential) solution to these issues is to consider effects both at the item level and the participant level through, e.g., linear mixed effects modelling. Predicting item success rather than aggregate performance (as was the case with the ACT in Chapter IV) allows for conclusions based on statistical approaches with more power. Although this approach could not have salvaged the test set of the Change Detection task, it is nonetheless an approach that better accounts for potential issues with particular items.

Although all the tasks presented here used accuracy, it is worth mentioning another method of measurement that was attempted: response time. A pilot version of the Swaps data suffered from ceiling effects (using only 3-letter and 2-step items), so we attempted to use response speed instead. In another study (not described in this thesis) aiming to measure implicit binding, response speed was used to gauge the effect of primed binding. Although speed worked better than accuracy for both of these studies, it was nonetheless difficult to interpret the scores. This is because we could not determine how much of the variance seen in speed was due to the hypothesized task effects or due to tendencies to take time with problems, to make recalculations, or to achieve perfection rather than satisfactory performance. Each of these tendencies are influenced by strategic differences (and related to speed-accuracy trade-offs) but the presented experiments were aiming to isolate specific cognitive processes associated with the manipulations. Although obviously strategy does account for variance in accuracy (as discussed in each chapter), the problems with response speed are that (a) it is harder to identify the influence of strategic differences (because it influences multiple ‘tendencies’, as described above) and (b) these strategic differences often have a far greater impact on the outcome of the data (in terms of systematic variance). For instance, whereas an accuracy error may be spotted by a vigilant and patient participant once in every 10 problems, the additional delay in response speed adds up on *every* trial. Appropriate difficulty was not always easy to get right (the initial version of the CD task was even harder before a pilot test, with performance bordering on chance in all conditions) but ultimately, perfect difficulty is not needed. It is only important that performance is above floor and below ceiling, with enough meaningful variance between individuals. A recurring fear for many of the studies was that the most difficult version of the task would be the one correlating with Gf, but we repeatedly observed this was unfounded. The DC version of the LST, the simple same matches in the RMT, and items of low Steps in the Swaps, were all among the best predictors of Gf in their respective experiments, despite also being among the easiest conditions in terms of accuracy.

Another measurement issue to consider was the potential for positive manifold to make it difficult to infer the relative importance of cognitive processes (Jensen, 1998), because it is impossible to truly strip a task of all “Gf-related” components. There are two reasons to be unconcerned with positive manifold affecting the interpretations. First, a pilot version of the RMT specifically aimed to address this by considering a ‘sensory discrimination’ condition where participants simply needed to identify one different number in the entire array. This was over potential concerns that any positive manifold was not just due to a general mental ability (Jensen, 1998) but due to *motivation* to engage in the tasks. Motivation was more concerning than general abilities because it could interact with the tasks being used. If motivation was carrying some of the correlations seen in Gf (i.e., more motivated participants are engaging in both Gf tasks and the predictor tasks), this sensory discrimination condition would help control for that by including it in the analyses, as it was assumed that the sensory discrimination required no ability and would thus only represent motivation. As it turned out, this condition had a virtually 0.0 correlation to Gf and was redundant for the full task that ended up in Chapter V. Thus, it seems there needs to be at least some threshold of task complexity to draw on resources in such a way that would lead to positive manifold. In the end though, these concerns were probably also unfounded. In several of the studies, a basic effect of positive manifold was already accounted for by the base task. For instance, the basic arithmetic in the control ACT had a significant correlation with Gf, though this was accounted for in the analyses by only considering what was unique to the experimental conditions. Thus, the second reason to be unconcerned was simply because we observed substantially different correlations both within tasks (e.g., between RMT conditions) and between tasks (e.g., between the RMT and complex-span), indicating that positive manifold – even it did exist – was not a problem for our approach as the specific processes in specific conditions and tasks were often isolated.

A final noteworthy contribution was the reliability (or lack thereof) of other, paradigmatic tasks. Despite the experimental modifications to the core task of each chapter, our approach throughout the thesis was to compare these modifications to established tasks to situate the conditions in the wider literature. Tasks included OSPAN, SSPAN, *n*-back, Letter Series, and, most commonly, Raven’s APM. Our mileage with these tasks varied greatly. Considering our aims (to situate our results in the wider literature), the most important aspect of these tasks was consistency. Raven’s APM was used often, partly because it is a well-respected measure in the wider literature (Carpenter et al., 1990), but also because it resoundingly achieved this goal of consistency. We repeatedly found a mean accuracy of around 60% with a standard deviation of around 20%, and a uniform distribution of scores with the upper range (scoring near perfectly) and lower range (scoring only a few items) seen in our university populations. Despite using a shortened 20-item version of the task, our distributions were remarkably consistent. The shortened version was also useful because it required an average of only 12 minutes to administer. The *n*-back also fared quite well with some strong correlations to expected tasks, though the scoring method of hits minus false alarms averaged over blocks meant the scale was not consistent with most other measures (proportion correct). The difficulty range on the *n*-back was also not as smooth as in Raven's APM or other WM measures, because each additional *n* caused such a substantial jump in difficulty (as opposed to difficulty increasing over 20 items in APM or from three to seven items in complex span). In general, we tended to use a 2-back set and a 3-back set because pilot testing indicated 1-back was too easy and 4-back too difficult though the lack of this range may have contributed to its more inconsistent correlations between studies. Nonetheless, the *n*-back tended to perform reasonably well. Our use of complex span tasks, comparatively, was frequently poor. Despite following the advice of prior research (Redick et al., 2012), the complex span paradigm consistently failed to meet the correlations observed elsewhere with both other WM tasks and APM. Although the lack of a correlation between *n*-back and complex span was expected, the failure to also correlate with the APM was somewhat concerning. Although this thesis has made arguments for why the complex span is unreliable (e.g., allows too much variation in chunking strategies), there was nonetheless some concern that we may have been administering the task wrong, given its frequent failings. However, the complex span did show hypothesized correlations with certain conditions (like ACT-retention and LST-Basic), they just tended to be weaker and more unreliable than expected. The occasional positive results indicate that it was employed correctly, but that it may simply not be the ideal WM task and Gf predictor that it is touted to be (Ackerman et al., 2005). Rather, it requires latent variable analysis with multiple versions (i.e., operation span, reading span, etc.) to smooth out these inconsistencies (discussed in more depth in the next section).

8.1.4. The future of Working Memory

A final point to consider is the possibilities for the future of this area of research. Working memory is an incredibly important topic to research. In lay understandings, it can be seen as ‘what we are thinking about at any one time’ or simply consciousness (Persuh, LaRock, & Berger, 2018). More advanced understandings extend that definition to include things recently thought of (with less activation) as well as the processes that enable the access and manipulation of these things. Given the importance of WM, it is no surprise that it is often a major topic of introductory psychology, often grouped in the broad term ‘memory’ (Griggs & Jackson, 2013). However, these topics typically limit discussion of WM to Baddeley’s model and (somewhat confusingly) separate WM from the discussion of chunking in short-term memory (e.g., Weiten, 2010). While the topic of WM has to be simplified for introductory psychology, the prolific nature of Baddeley’s model has gone far beyond textbooks and often represents psychologists’ only interpretation of WM (for those outside of WM research).[24]

Often, researchers outside WM that need to employ a WM task rely on Baddeley’s model and the task that was designed to simultaneously tap the two components of Baddeley’s model: the complex span paradigm. Engle’s (2002) addition of executive attention into the WM understanding proved seminal (with over 2,000 citations); probably because it addressed much of the insufficiencies over the central executive in Baddeley’s model. For researchers looking for a quick understanding of WM, frustration over the ambiguity of central executive would have been met with Baddeley’s (2003) admission that it

[24] A frequent challenge in presenting this research to other psychologists (e.g., at conferences) was overcoming the view that working memory is primarily a storage system with a central executive.

is indeed ambiguous and largely a placeholder. For these researchers, a brief, simple, and well-written review (Engle, 2002) on a very human construct that intuitively should be related to WM (executive attention) was a perfect replacement citation. Particularly so, because it still advocated for the use of complex span paradigm, presenting research in all sorts of modalities (reading, operation) with the same basic premise: store information while processing other information. As discussed (and observed) throughout this thesis, the complex span appears inadequate for measuring WM, at least in isolation. It can lead to similar conclusions as simple spans (Colom, Rebollo, et al., 2006), despite missing the component that makes it integrally WM. The processing components across the specific tasks (e.g., symmetry vs. operation) are not made equally (i.e., symmetry judgements vs. mathematical operations) and the correlations between the variations of the tasks are not particularly strong (Conway et al., 2005). In all, complex span tasks in themselves are quite impure measures of WM and the way to resolve this impurity is to use latent variable analysis with multiple complex span tasks, purifying the tasks by extracting what is shared (Conway et al., 2005). In this way, the complex span paradigm is not just the general task procedure of ‘storage with intermittent processing’; but rather, the paradigm must necessarily include the latent variable extraction performed on *multiple* tasks with the ‘storage with intermittent processing’ procedure. Admittedly, our use of the complex span paradigm throughout this thesis lacked this latent extraction component, but for good reason (see Section 1.3.3.2). What was unexpected was that the complex span *required* latent extraction to perform as a measure of WM at all. The task analyses performed throughout this thesis have hopefully provided good reasons to consider cognitive processes at the task level, where task administration is pragmatic (requiring shorter testing times and potential reductions in fatigue) and the experimental manipulations can be directly related to cognitive demands, rather than related through shared processes across tasks. Although either approach is valid (*with* theoretical accounts of the disparate/shared processes), the risk for the future of working memory is the interpretation of the shared processes as coming from established models (attentional control, multicomponential models) rather than from new, falsifiable models. For instance, Engle's (2018) process-oriented account of more attention-focused tasks (like Stroop and anti-saccade) is much more compelling (than complex span approaches), with these tasks often being the most reliable among their arsenal.

The complex span paradigm may indeed be a valid and reliable approach (*with* the necessary inclusion of latent variable analysis) but, as Conway et al. (2005) state, it should not represent the 'gold standard' of WM measures. The risk with a gold standard complex span, when coupled with a pervasive (i.e., widely understood) theory such as the multicomponential model (Baddeley & Hitch, 1974), is a risk of continued inertia in the field. Lay understandings of WM will continue to be difficult to reconcile with chunking theories in short-term memory, which in turn, make the leaps towards analogical reasoning accounts (on the higher end) and visual encoding accounts (on the lower end) harder than they need to be, despite both ends of the spectrum being so closely related to a foundation of relational integration in working memory. The current thesis has hopefully made these leaps more tangible, with discussion on analogical reasoning (throughout) and the lower-level accounts like visual WM (in Chapter VII) assimilated with the more balanced perspective of a cognitive-relational WM. The integration of somewhat distant approaches in the experimental-differential approach (experimental manipulations and individual differences) used here has led to insightful and powerful findings on working memory. A future for working memory may well depend on continued integration of disparate approaches (like analogical reasoning and visual WM) along with more process-oriented accounts of task performance, lest we continue to offer overly dominant but potentially confusing explanations, such as the central executive.

8.2. Conclusion

Overall, the goal of this project was to demonstrate how working memory can be understood through relational integration. We have discussed how established tasks can have their loads conceptualized as binding capacity required for relational integration, rather than storage capacity required simply for keeping information active. Across five studies, we have demonstrated that tasks can be designed to place a premium on binding and unbinding processes required to integrate a connected relation. We have seen some evidence that this is the core of working memory across functioning humans (e.g., the change detection task which put a limit on performance with the design of items) but also that individuals can vary greatly in their relational integration ability, and that this may be the purest link to more complex constructs like fluid intelligence. While the ability to process relations and interpret analogies has no doubt been recognised as an important aspect of human cognition, the field of working memory has struggled to see past modular views separating storage, attention, and processing, particularly in operationalization of demands. While the gold standard for the field has been to use latent variable analysis to correlate already established tasks, the studies throughout this thesis demonstrate the potential of task analyses leading to experimental manipulations that can infer powerful judgements about the demands of working memory.

REFERENCES

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working Memory and Intelligence: The Same or Different Constructs? *Psychological Bulletin, 131*(1), 30-60.

Alexander, P. A. (2016). Relational thinking and relational reasoning: Harnessing the power of patterning. *NPJ Science of Learning, 1*, 7.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*, 261-295.

Atkinson, R. C., & Shiffrin, R. M. (1968). Chapter: Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89-195). New York: Academic Press.

Baddeley, A. D. (1992). Working Memory. *Science, 255*(5044), 556-559.

Baddeley, A. D. (1993). Working memory or working attention? In A. D. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control* (pp. 152-170). Oxford: Clarendon Press.

Baddeley, A. D. (1998). The central executive: A concept and some misconceptions. *Journal of the International Neuropsychological Society, 4*, 523-526.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417-423.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*, 829-839.

Baddeley, A. D. (2012). Working memory: Theories, models and controversies. *Annual Review of Psychology, 63*, 1-29.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47-89). New York: Academic Press.

Bateman, J. E. (2015). *Exploring the components of working memory related to fluid intelligence through integration and coordination of storage and processing load.* (Bachelor of Psychology, Honours), University of Sydney, Australia.

Bateman, J. E., & Birney, D. P. (2019). The link between working memory and fluid intelligence is dependent on flexible bindings, not passive or systematic retention. *Acta Psychologica, 199*, Advanced online publication.

Bateman, J. E., Birney, D. P., & Loh, V. (2017). *Exploring functions of working memory related to fluid intelligence: Coordination, relational integration, and access*. Paper presented at the 39th Annual Meeting of the Cognitive Science Society, London, UK.

Bateman, J. E., Ngiam, W. X. Q., & Birney, D. P. (2018). Relational encoding of objects in working memory: Change detection performance is better for violations in group relations. *PLoS ONE, 13*(9).

Bateman, J. E., Thompson, K. A., & Birney, D. P. (2019). Validating the relation monitoring task as a measure of relational integration and predictor of fluid intelligence. *Memory & Cognition*, *47*(8), 1457-1468. doi:https://doi.org/10.3758/s13421-019-00952-2

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2017). lme4: Linear Mixed-Effects Models using 'Eigen' and S4. https://cran.r-project.org/web/packages/lme4/lme4.pdf. .

Birney, D. P. (2002). *The measurement of task complexity and cognitive ability: Relational complexity in adult reasoning.* (PhD), University of Queensland, Australia.

Birney, D. P. (n.d.-a). *Difficulty of the Swaps Task determined by distance of swaps*. Unpublished raw data.

Birney, D. P. (n.d.-b). *Verbal protocol analysis of the Latin Square Task*. Unpublished data.

Birney, D. P., Beckmann, J. F., & Beckmann, N. (2019). Within-individual variability of ability and learning trajectories in complex problems. In D. J. McFarland (Ed.), *General and Specific Mental Abilities* (pp. 253-283). Cambridge, UK: Cambridge Scholars Publishing.

Birney, D. P., & Bowman, D. B. (2009). An experimental-differential investigation of cognitive complexity. *Psychology Science Quarterly, 51*(4), 449-469.

Birney, D. P., Bowman, D. B., Beckmann, J. F., & Seah, Y. Z. (2012). Assessment of processing capacity: Reasoning in Latin Square Tasks in a population of managers. *European Journal of Psychological Assessment, 28*(3), 216-226.

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*(1), 146-171.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences, 29,* 109-160.

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology, 49*, 229-240.

Bower, G. H. (1970). Analysis of a mnemonic device: Modern psychology uncovers the powerful components of an ancient system for improving memory. *American Scientist, 58*(5), 496-510.

Bowman, D. B. (2006). *An investigation of the determinants of cognitive complexity and individual differences in fluid reasoning ability.* (PhD), University of Sydney, Australia.

Brady, T. F., & Alvarez, G. A. (2014). No evidence for a fixed object limit in working memory: Spatial ensemble representations inflate estimates of working memory capacity for complex objects. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 41*(3), 921-929.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review, 120*(1), 85-109.

Broadbent, D. E. (1958). *Perception and communication*. New York, NY: Pergamon Press.

Buehner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2006). Cognitive abilities and their interplay: Reasoning, crystallized intelligence, working memory components, and sustained attention. *Journal of Individual Differences, 27*(2), 57-72.

Bui, M., & Birney, D. P. (2014). Learning and individual differences in Gf processes and Raven's. *Learning and Individual Differences, 32,* 104-113.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What on intelligence test measures: A threoetical account of the processing in the Raven Progressive Matrices test. *Psychological Review, 97*(3), 404-431.

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15*(3), 139-164.

Chekaf, M., Gauvrit, N., Guida, A., & Mathy, F. (2015). *Chunking in working memory and its relationship to intelligence*. Paper presented at the 37th Annual Meeting of the Cognitive Science Society, Pasadena, California.

Chuderski, A. (2014). The relational integration task explains fluid reasoning above and beyond other working memory tasks. *Memory & Cognition, 42*, 448-463.

Chuderski, A. (2015). The broad factor of working memory is virtually isomorphic to fluid intelligence tested under time pressure. *Personality and Individual Differences, 85,* 98-104.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Ed. ed.). Hillsdale, NJ: Erlbaum.

Cohen, N. J., & Eichenbaum, H. B. (1993). *Memory, amnesia, and the hippocampal system*. London: The MIT Press.

Colflesh, G. J. H., & Conway, A. R. A. (2007). Individual differences in working memory capacity and divided attention in dichotic listening. *Psychonomic Bulletin & Review, 14*(4), 699-703.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*(6), 407-428.

Colom, R., Rebollo, I., Abad, F. J., & Shih, P. C. (2006). Complex span tasks, simple span tasks, and cognitive abilities: A reanalysis of key studies. *Memory & Cognition, 34*(1), 158-171.

Colom, R., Shih, P. C., Flores-Mendoza, C., & Quiroga, Á. M. (2006). The real relationship between short-term memory and working memory. *Memory, 14*(7), 804-813.

Conlin, J. A., Gathercole, S. E., & Adams, J. W. (2005). Children's working memory: Investigating performance limitations in complex span tasks. *Journal of Experimental Child Psychology, 90*(4), 303-317.

Conway, A. R. A., Cowan, N., & Bunting, M. F. (2001). The cocktail party phenomenon revisited: The importance of working memory capacity. *Psychonomic Bulletin & Review, 8*(2), 331-335.

Conway, A. R. A., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. B. (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence, 30*, 163-183.

Conway, A. R. A., Jarrold, C., Kane, M. J., Miyake, A., & Towse, J. (2007). *Variation in working memory* (A. R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. Towse Eds.). Oxford, UK: Oxford University Press.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review, 12*(5), 769-786.

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104*(2), 163-191.

Cowan, N. (1995). *Attention and memory: An integrated framework* (Vol. No. 265). New York: Oxford University Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-185.

Cowan, N. (2005). *Working Memory Capacity*. Hove, UK: Psychology Press.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science, 19*(1), 51-57.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*, 1158-1170.

Cowan, N., Winkler, I., Teder, W., & Näätänen, R. (1993). Memory prerequisites of mismatch negativity in the auditory event-related potential (ERP). *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19*(4), 909-921.

Crawford, J. D. (1988). *Intelligence, task complexity and tests of sustained attention.* (PhD), The University of New South Wales,

Crowder, R. G. (1979). Similarity and order in memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 13). New York: Academic Press.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual Review of Psychology, 66*, 115-142.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*, 450-466.

Deary, I. J. (2001). Human intelligence differences: towards a combined experimental-differential approach. *Trends in Cognitive Sciences, 5*(4), 164-170.

Deary, I. J. (2003). Ten things I hated about intelligence research. *The Psychologist, 16*(10), 534-537.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition, 44*, 1-42.

Dent, K. (2009). Coding categorical and coordinate spatial relations in visual-spatial short-term memory. *The Quaterly Journal of Experimental Psychology, 62*(12), 2372-2387.

DeStefano, D., & LeFevre, J.-A. (2004). The role of working memory in mental arithmetic. *European Journal of Cognitive Psychology, 16*(3), 353-386.

Detterman, D. K. (1989). The future of intelligence research. *Intelligence, 13*, 199-203.

Dilevski, N. (2016). *Unraveling the relationship between complex span tasks and the n-back task of working memory: An experimental-differential approach.* (Bachelor of Psychology Honours), University of Sydney, Australia.

Donald, M. (1991). *Origins of the modern mind: Three stages in the evolution of culture and cognition*. Cambridge, MA: Harvard University Press.

Dumas, D., Alexander, P. A., & Grossnickle, E. M. (2013). Relational reasoning and its manifestation in the educational context: A systematic review of the literature. *Educational Psychology Review, 25*, 391-427.

Engle, R. W. (1996). Working memory and retrieval: An inhibition-resource approach. In J. Richardson, R. W. Engle, L. Hasher, R. H. Logie, E. Stolzfus, & R. Zacks (Eds.), *Working memory and human cognition* (pp. 89-119). New York: Oxford University Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*(1), 19-23.

Engle, R. W. (2018). Working memory and executive attention: A revisit. *Perspectives on Psychological Science, 12*(2), 190-193.

Engle, R. W., Cantor, J., & Carullo, J. J. (1992). Individual differences in working memory and comprehension: A test of four hypotheses. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 18*(5), 972-992.

Engle, R. W., & Kane, M. J. (2004). Executive attention, working memory capacity, and a two-factor theory of cognitive control. In B. Ross (Ed.), *The Psychology of Learning and Motivation* (Vol. 44, pp. 145-199). New York: Elsevier.

Engle, R. W., Kane, M. J., & Tuholiski, S. W. (1999). Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. In A. Miyake & P. Shah (Eds.), *Models of working memory*. Cambridge, UK: Cambridge University Press.

Engle, R. W., Tuholiski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309-331.

Feigenson, L., & Halberda, J. (2008). *Conceptual knowledge increases infants' memory capacity.* Paper presented at the Proceedings of the National Academy of Sciences of the United States of America.

Fougnie, D. (2008). The relationship between attention and working memory. In N. B. Johansen (Ed.), *New Research on Short-Term Memory* (pp. 1-45). New York: Nova Science Publishers.

Gabales, L., & Birney, D. P. (2011). Are the limits in processing and storage capacity common? Exploring the additive and interactive effects of processing and storage load in working memory. *Journal of Cognitive Psychology, 23*(3), 322-341.

Garcia, J. (2008). *Cartesian Theatre* [Image]. In. Retrieved from https://commons.wikimedia.org/wiki/File:Cartesian_Theater.jpg.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D. (2003). Why we're so smart. In D. Gentner & S. Goldin-Meadow (Eds.), *Language in mind: Advances in the study of language and thought* (pp. 195-235). Cambridge, MA: MIT Press.

Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual differences research. *Personality and Individual Differences, 102*, 74-78.

Gimmig, D., Huguet, P., Caverni, J.-P., & Cury, F. (2006). Choking under pressure and working memory capacity: When performance pressures reduces fluid intelligence. *Psychonomic Bulletin & Review, 13*(6), 1005-1010.

Giofrè, D., Borella, E., & Mammarella, I. C. (2017). The relationship between intelligence, working memory, academic self-esteem, and academic achievement. *Journal of Cognitive Psychology, 29*(6), 731-747.

Griggs, R. A., & Jackson, S. L. (2013). Introductory psychology textbooks: An objective analysis update. *Teaching of Psychology, 40*(3), 163-168.

Halford, G. S., Baker, R., McCredden, J. E., & Bain, J. D. (2005). How many variables can humans process? *Psychological Science, 16*(1), 70-76.

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

Hasher, L., & Zacks, R. T. (1988). Working memory, comprehension, and aging: A review and a new view. *Psychology of Learning and Motivation, 22*, 193-225.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270.

Hummel, J. E., & Holyoak, K. J. (2001). A process model of human transitive inference. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 279-305). Cambridge, MA: The MIT Press.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*(2), 220-264.

iMotions Biometric Research Platform. (2018). iMotions Version 7.1 [Computer Program] (Version 7.1). Copenhagen, Denmark: iMotions A/S.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences of the United States of America, 105*(19), 6829-6833.

Jensen, A. R. (1977). The nature of intelligence and its relation to learning. *Melbourne Studies in Education, 20*(1), 107-133.

Jensen, A. R. (1980). *Bias in mental testing*. New York, N. Y.: The Free Press.

Jensen, A. R. (1998). *The g factor: The science of mental ability*. London: Praeger.

Jiang, Y., Chun, M. M., & Olson, I. R. (2004). Perceptual grouping in change detection. *Perception & Psychophysics, 66*(3), 446-453.

Jiang, Y., Olson, I. R., & Chun, M. M. (2000). Organization of visual short-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 26*(3), 683-702.

Kane, M. J., Bleckley, M. K., Conway, A. R. A., & Engle, R. W. (2001). A controlled-attention view of working memory capacity. *Journal of Experimental Psychology: General, 130*(2), 169-183.

Kane, M. J., & Engle, R. W. (2003). Working memory capacity and the control of attention: The contributions of goal neglect, response competition, and task set to Stroop interference. *Journal of Experimental Psychology: General, 132*(1), 47-70.

Kane, M. J., Hambrick, D. Z., Tuholiski, S. W., Wilhelm, O., Payne, T., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology: General, 133*, 189-217.

Kessler, Y., & Oberauer, K. (2014). Working memory updating latency reflects the cost of switching between maintenance and updating modes of operation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 40*(3), 738-754.

Koenig, K. A., Frey, M. C., & Detterman, D. K. (2008). ACT and general cognitive ability. *Intelligence, 36*, 153-160.

Krumm, S., Lipnevich, A. A., Schmidt-Atzert, L., & Bühner, M. (2012). Relational integration as a predictor of academic achievement. *Learning and Individual Differences, 22*, 759-769.

Krumm, S., Schmidt-Atzert, L., Buehner, M., Ziegler, M., Michalczyk, K., & Arrow, K. (2009). Storage and non-storage components of working memory predicting reasoning: A simultaneous examination of a wider range of ability factors. *Intelligence, 37*, 347-364.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity. *Intelligence, 14*, 389-433.

Laidra, K., Pullmann, H., & Allik, J. (2007). Personality and intelligence as predictors of academic achievement. *Personality and Individual Differences, 42*, 441-451.

Laurence, P. G., Mecca, T., P., Serpa, A., Martin, R., & Macedo, E. C. (2018). Eye movements and cognitive strategy in a fluid intelligence test: Item type analysis. *Frontiers in Psychology, 9*, 9.

Lewandowsky, S., Geiger, S. M., Morrell, D. B., & Oberauer, K. (2010). Turning simple span into complex span: Time for decay or interference from distractors? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(4), 958-978.

Li, K. Z. H. (1999). Selection from working memory: On the relationship between processing and storage components. *Aging, Neuropsychology, and Cognition, 6*(2), 99-116.

Lilienthal, L., Tamez, E., Myerson, J., & Hale, S. (2013). Predicting performance on the Raven's Matrices: The roles of associative learning and retrieval efficiency. *Journal of Cognitive Psychology, 25*(6), 704-716.

Loesche, P., Wiley, J., & Hasselhorn, M. (2015). How knowing the rules affects solving the Raven Advanced Progressive Matrices Test. *Intelligence, 48*, 58-75.

Logie, R. H. (1996). The seven ages of working memory. In J. T. E. Richardson, R. W. Engle, L. Hasher, R. H. Logie, E. R. Stoltzfus, & R. T. Zacks (Eds.), *Working Memory and Human Cognition* (pp. 31-65). New York: Oxford University Press.

Lohman, D. F., & Lakin, J. M. (2011). Intelligence and reasoning. In R. J. Sternberg & S. B. Kaufman (Eds.), *The Cambridge Handbook of Intelligence* (pp. 419-441). Cambridge, UK: Cambridge University Press.

Lu, J., Tian, L., Zhang, J., Wang, J., Ye, C., & Liu, Q. (2017). Strategic inhibition of distractors with visual working memory contents after involuntary attention capture. *Scientific Reports, 7*.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature, 390*, 279-281.

Lüdecke, D. (2017). sjPlot: Data Visualization for Statistics in Social Science (Version 2.0) [Computer software]. Retrieved from http://CRAN.R-project.org/package=sjPlo

Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: A user's guide*. Cambridge: Cambridge University Press.

Manning, J. R., & Kahana, M. J. (2012). Interpreting semantic clustering effects in free recall. *Memory, 20*(5), 511-517.

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills, 112*(2), 331-348.

Matešić, K. (2000). Relations between results on Raven progressive matrices plus sets and school achievement. *Review of Psychology, 7*(1-2), 75-82.

Mathy, F., Chekaf, M., & Cowan, N. (2018). Simple and complex working memory tasks allow similar benefits of information compression. *Journal of Cognition, 1*(1), 1-12.

McLean, R. S., & Gregg, L. W. (1967). Effects of induced chunking on temporal aspects of serial recitation. *Journal of Experimental Psychology, 74*(4), 455-459.

Mervyn, S., Gewer, A., Osrin, Y., Khunou, D., Fridjhon, P., & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's matrices scores of African and non-African university students in South Africa. *Intelligence, 30*, 221-232.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Millisecond Software. (2014). Inquisit 4.0.8.0 Version 2624. In. Seattle, WA: Millisecond Software.

Millisecond Software. (2017). Inquisit Lab 5 [Computer Software] (Version 5.0.8.0). Seattle, WA: Millisecond Software.

Miyake, A., & Shah, P. (1999). Models of working memory: An introduction. In *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge: Camridge University Press.

Murdock Jr., B. B. (1966). Visual and auditory stores in short-term memory. *The Quarterly Journal of Experimental Psychology, 18*(3), 206-211.

Oberauer, K. (2002). Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 28*(3), 411-421.

Oberauer, K. (2009a). Design for a working memory. In *Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 51, pp. 45-100). San Diego: Elsevier Academic Press Inc.

Oberauer, K. (2009b). Interference between storage and processing in working memory: Feature overwriting, not similarity-based competition. *Memory & Cognition, 37*(3), 346-357.

Oberauer, K., Awh, E., & Sutterer, D. W. (2017). The role of long-term memory in a test of visual working memory: proactive facilitation but no proactive interference. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 43*(1), 1-22.

Oberauer, K., Demmrich, A., Mayr, U., & Kliegl, R. (2001). Dissociating retention and access in working memory: An age-comparative study of mental arithmetic. *Memory & Cognition, 29*(1), 18-33.

Oberauer, K., Farrell, S., Jarrold, C., Pasiecznik, K., & Greaves, M. (2012). Interference between maintenance and processing in working memory: The effect of item-distractor similarity in complex span. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*(3), 665-685.

Oberauer, K., & Hein, L. (2012). Attention to information in working memory. *Current Directions in Psychological Science, 21*(3), 164-169.

Oberauer, K., & Lewandowsky, S. (2008). Forgetting in immediate serial recall: Decay, temporal distinctiveness, or interference? *Psychological Review, 115*(3), 544-576.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A. R. A., Cowan, N., . . . Ward, G. (2018). Benchmarks for models of short term and working memory. *Psychological Bulletin*.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. *Psychonomic Bulletin & Review, 19*, 779-819.

Oberauer, K., Süß, H. M., Wilhelm, O., & Sander, N. (2007). Individual differences in working memory capacity and reasoning ability. In R. A. Conway, C. Jarrold, M. J. Kane, A. Miyake, & J. N. Yowse (Eds.), *Variation in Working Memory*. New York: Oxford University Press.

Oberauer, K., Süß, H. M., Wilhelm, O., & Wittman, W. W. (2003). The multiple faces of working memory: Storage, processing, supervision, and coordination. *Intelligence, 31*, 167-193.

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

Olsen, R. K., Lee, Y., Kube, J., Rosenbaum, R. S., Grady, C. L., Moscovitch, M., & Ryan, J. D. (2015). The role of relational binding in item memory: Evidence from face recognition in a case of developmental amnesia. *The Journal of Neuroscience, 35*(13), 5342-5350.

Olson, I. R., & Newcombe, N. S. (2014). Binding together the elements of episodes: Relational memory and the developmental trajectory of the hippocampus. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell Handbook on the Development of Children's Memory* (pp. 285-308). West Sussex, United Kingdom: Wiley-Blackwell.

Parkin, A. J. (1998). The central executive does not exist. *Journal of the International Neuropsychological Society, 4*, 518-522.

Paul, S. T., Monda, S., Olausson, S. M., & Reed-Daley, B. (2014). Effects of apophenia on multiple-choice exam performance. *SAGE Open, 4*(4). doi:10.1177/2158244014556628

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences, 31*, 109-178.

Perret, P., Bailleux, C., & Dauvier, B. (2011). The influence of relational complexity and strategy selection on children's reasoning in the Latin Square Task. *Cognitive Development, 26*(2), 127-141.

Persuh, M., LaRock, E., & Berger, J. (2018). Working memory and consciousness: The current state of play. *Frontiers in Human Heuroscience, 12*, 1-7. doi:10.3389/fnhum.2018.00078

Poirier, M., & Saint-Aubin, J. (1995). Memory for related and unrelated words: Further evidence on the influence of semantic factors in immediate serial recall. *Quarterly Journal of Experimental Psychology, 48*(2), 384-404.

Portrat, S., Guida, A., Phénix, T., & Lemaire, B. (2016). Promoting the experimental dialogue between working memory and chunking: Behavioral data and simulation. *Memory & Cognition, 44*(3), 420-434.

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.0). Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., Spybrook, J., Bloom, H., Congdon, R., Hill, C., & Martinez, A. (2011). Optimal Design Software for Multi-level and Longitudinal Research (Version 3.01). In: Available from www.wtgrantfoundation.org

Raven, J. (1989). The Raven Progressive Matrices: A review of national norming studies and ethnic and socioeconomic variation within the United States. *Journal of Educational Measurement, 26*, 1-16.

Raven, J. C. (1941). Standardisation of progressive matrices. *British Journal of Psychology, XIX*(1), 137-150.

Redick, T. S., Broadway, J. M., Meier, M. E., Kuriakose, P. S., Unsworth, N., Kane, M. J., & Engle, R. W. (2012). Measuring working memory capacity with automated complex span tasks. *European Journal of Psychological Assessment, 28*, 164-171.

Redick, T. S., & Lindsey, D. R. B. (2013). Complex span and *n*-back measures of working memory: A meta-analysis. *Psychonomic Bulletin & Review, 20*(3), 1102-1113.

Rensink, R. A. (2000a). Four-sight in hindsight: The existence of magical numbers in vision [Peer commentary on “The magical number 4 in short-term memory: A reconsideration of mental storage capacity” by N. Cowan]. *Behavioral and Brain Sciences, 24,* 87-185.

Rensink, R. A. (2000b). Visual search for change: A probe into the nature of attentional processing. *Visual Cognition, 7*(1/2/3), 345-376.

Rensink, R. A. (2002). Change detection. *Annual Review of Psychology, 53,* 245-277.

Resnick, I., Davatzes, A., Newcombe, N. S., & Shipley, T. F. (2016). Using relational reasoning to learn about scientific phenomena at unfamiliar scales. *Educational Psychology Review*, 1-15. doi:10.1007/s10648-016-9371-5

Revelle, W. (2018). psych: Procedures for personality and psychological research (Version 1.8.12). Northwestern University, Illinois, USA.

Richardson, J. T. E. (2007). Measures of short-term memory: A historical review. *Cortex, 43*(5), 635-650.

Ridgeway, D. (2006). Strategic grouping in the spatial span memory task. *Memory, 14*(8), 990-1000.

Robertson, I., Gratton, L., & Sharpley, D. (1987). The psychometric properties and design of managerial assessment centres: Dimensions into exercises won't go. *Journal of Occupational Psychology, 60,* 187-195.

Robin, N., & Holyoak, K. J. (1995). Relational complexity and the functions of prefrontal cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences*. MA: MIT Press.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*(4), 803-814.

Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral and Brain Sciences, 26*, 709-777.

Saint-Aubin, J., Ouellette, D., & Poirier, M. (2005). Semantic similarity and immediate serial recall: Is there an effect on all trials? *Psychonomic Bulletin & Review, 12*(1), 171-177.

Saint-Aubin, J., & Poirier, M. (1999). Semantic similarity and immediate serial recall: Is there a detrimental effect on order information? *The Quarterly Journal of Experimental Psychology, 52*(2), 367-394.

Schmidt, F. L., & Hunter, J. E. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173.

Shipstead, Z., Harrison, T. L., & Engle, R. W. (2016). Working memory capacity and fluid intelligence: Maintenance and disengagement. *Perspectives on Psychological Science, 11*(6), 771-799.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359-1366.

Sluzenski, J., Newcombe, N. S., & Kovacs, S. L. (2006). Binding, relational memory, and recall of naturalisitc events: A developmental perspective. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 32*(1), 89-100.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied, 74*(11), 1-29.

Squire, L. R. (2004). Memory systems of the brain: A brief history and current perspective. *Neurobiology of Learning and Memory, 82*, 171-177.

Stankov, L. (2000). Complexity, metacognition, and fluid intelligence. *Intelligence, 28*(2), 121-143.

Stankov, L., & Crawford, J. D. (1993). Ingredients of complexity in fluid intelligence. *Learning and Individual Differences, 5*(2), 73-111.

Stankov, L., & Schweizer, K. (2007). Raven's Progressive Matrices, manipulations of complexity and measures of accuracy, speed and confidence. *Psychology Science, 49*(4), 362-342.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84*(4), 353-378.

Tan, L., & Ward, G. (2008). Rehearsal in immediate serial recall. *Psychonomic Bulletin & Review, 15*(3), 535-542.

Thompson, V. A., Prowse Turner, J. A., & Pennycook, G. (2011). Intuition, reason, and metacognition. *Cognitive Psychology, 63*(3), 107-140.

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Unsworth, N., & Engle, R. W. (2005). Working memory capacity and fluid abilities: Examining the correlation between Operation Span and Raven. *Intelligence, 33*, 67-81.

Unsworth, N., & Engle, R. W. (2006). Simple and complex memory spans and their relation to fluid abilities: Evidence from list-length effects. *Journal of Memory and Language, 54*, 68-80.

Unsworth, N., & Engle, R. W. (2007a). The nature of individual differences in working memory capacity: Active maintenance in primary memory and controlled search from secondary memory. *Psychological Review, 114*(1), 104-132.

Unsworth, N., & Engle, R. W. (2007b). On the division of short-term and working memory: An examination of simple and complex span and their relation to higher order abilities. *Psychological Bulletin, 133*(6), 1038-1066.

Unsworth, N., Spillers, G. J., & Brewer, G. A. (2010). The contribution of primary and secondary memory to working memory capacity: An individual differences analysis of immediate free recall. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(1), 240-247.

van der Linden, M. (1998). The relationships between working memory and long-term memory. *Comptes rendus de l'Académie des Sciences, 321*(2-3), 175-177.

van Lamsweerde, A. E., Beck, M. R., & Johnson, J. S. (2016). Visual working memory organization is subject to top-down control. *Psychonomic Bulletin & Review, 23*(4), 1181-1189.

Vendetti, M. S., Wu, A., & Holyoak, K. J. (2014). Far-out thinking: Generating solutions to distant analogies promotes relational thinking. *Psychological Science, 25*(4), 928-933.

Verguts, T., & De Boeck, P. (2002). The induction of solution rules in Raven's Progressive Matrices Test. *European Journal of Cognitive Psychology, 14*(4), 521-547.

Vidal, J. R., Gauchou, H. L., Tallon-Baudry, C., & O'Regan, K. (2005). Relational information in visual short-term memory: The structural gist. *Journal of Vision, 5*(3), 244-256.

Vogel, E. K., Woodman, G. F., & Luck, S. J. (2006). The time course of consolidation in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(6), 1436-1451.

Waterstone, P. X. (2007). Apophenia & Illusory Correlation. Retrieved from https://waterstone.wordpress.com/2007/05/24/apophenia-illusory-correlation/

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale - Four Edition*. San Antonio, Texas: Pearson.

Weiten, W. (2010). *Psychology: Themes and variations* (Vol. 8). Belmont, CA: Wadsworth.

Wicherts, J. M., Dolan, C. V., Carlson, J. S., & van der Maas, H. L. J. (2010). Raven's test performance of sub-Saharan Africans: Average performance, psychometrics properties, and the Flynn effect. *Learning and Individual Differences*, *20*(3), 135-151.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag.

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity and how can we measure it? *Frontiers in Psychology, 4*. doi:10.3389/fpsyg.2013.00433

Woodman, G. F., Vecera, S. P., & Luck, S. J. (2003). Perceptual organization influences visual working memory. *Psychonomic Bulletin & Review*, *10*(1), 80-87.

Zeuch, N., Holling, H., & Kuhn, J.-T. (2011). Analysis of the Latin Square Task with linear logistic test models. *Learning and Individual Differences, 21*, 629-632.

APPENDIX A

The following analyses are the outcome of the regression models run in Section 3.9.4 using the ‘SSPAN-included’ approach. This lowers the sample size to *n* = 80 and concerns over using this subsample are described in Section 3.9.2. In the interest of transparency, the results of this approach are nonetheless given below, in an identical format as to how it would have appeared in the chapter, including the text.

We controlled for WM by adding SSPAN, OSPAN, and *n*-back to a preliminary model, which on its own predicted the *Gf* factor ($R^2 = .406$, $p < .001$), with SSPAN and *n*-back providing a unique contribution, but not OSPAN (SSPAN $sr^2 = .09$, $p = .001$; OSPAN $sr^2 < .01$, $p = .856$; *n*-back $sr^2 = .10$, $p = .001$). Adding LST-Basic was a significant increase ($\Delta R^2 = .031$, $p = .047$) though adding LST-DC on top of this was not ($\Delta R^2 = .017$, $p = .142$). Replicating this regression predicting just APM (rather than the *Gf* factor) resulted in a largely identical pattern of results *except* that LST-DC was a significant, unique contributor in the final model ($\Delta R^2 = .042$, $p = .016$). These two regressions, one predicting *Gf* and one predicting APM, are presented in Table A1 and Table A2, respectively. Running the regression predicting *Gf* without LST-Basic once again demonstrated that LST-Basic and LST-DC were largely contributing the same variance, as LST-DC became a significant change on its own, above the three WM measures ($\Delta R^2 = .043$, $p = .020$).

Table A1. *Full Regression Model Predicting the Gf factor in Experiment 3 using the SSPAN-included sample only.*

Model	Predictor	B	*t*	*p*	*sr²*	R²	ΔR²
1	Operation Span	.003	0.18	.856	< .001	.406	**.406**
	Symmetry Span	**.054**	3.32	001	.088		
	n-back	**.210**	3.54	.001	.101		
2	Operation Span	.002	0.09	.927	< .001	.438	**.031**
	Symmetry Span	**.054**	3.36	.001	.087		
	n-back	**.170**	2.77	.007	.059		
	LST-Basic	**.078**	2.02	.047	.031		
3	Operation Span	.006	0.32	.751	.001	.454	.017
	Symmetry Span	**.055**	3.42	.001	.089		
	n-back	**.154**	2.50	.015	.047		
	LST-Basic	.034	0.83	.410	.005		
	LST-DC	.073	1.45	.142	.017		

N=77; bold coefficients p < .05.

Table A2. *Full Regression Model Predicting APM in Experiment 3 using the SSPAN-included sample only.*

Model	Predictor	B	*t*	*p*	*sr²*	R²	ΔR²
1	Operation Span	-.036	-0.42	.678	.001	.431	**.431**
	Symmetry Span	**.297**	3.76	< .001	.109		
	n-back	**1.036**	3.62	.001	.101		
2	Operation Span	-.044	-0.52	.607	.002	.461	**.030**
	Symmetry Span	**.295**	3.81	< .001	.107		
	n-back	**.844**	2.85	.006	.060		
	LST-Basic	**.375**	2.02	.047	.030		
3	Operation Span	-.013	-0.15	.880	< .001	.503	**.042**
	Symmetry Span	**.299**	4.00	< .001	.110		
	n-back	**.721**	2.48	.015	.043		
	LST-Basic	.069	0.32	.753	.001		
	LST-DC	**.571**	2.47	.016	.042		

N=77; bold coefficients p < .05.

APPENDIX B

Table B1 demonstrates the first analysis conducted (Table B2 indicates contrast coding) in the ACT study from Chapter IV using *n* = 64. Note that these results use standard (disregarding recall accuracy) scoring rather than absolute scoring.

Table B1. *Model 1 and Model 2 Fixed and Random Effects Estimates*

				Fixed Effects		**Random Effects**
Predictors	**Model**	**OR**	**se**	**CI**	**p**	**tau**
(Intercept)	1	3.72	0.14	2.84 – 4.87	**<0.001**	0.78
Retention (vs Control) Cost	1	0.99	0.27	0.58 – 1.69	0.978	0.44
Access (vs Retention) Cost	2	0.26	0.23	0.16 – 0.40	**<0.001**	0.54
Binding (Fixed vs Random) Cost	1	0.66	0.21	0.44 – 0.99	**0.046**	0.81
RB moderator	1	1.47	0.14	1.11 – 1.94	**0.008**	0.09
Retention (vs Control) Cost x RB	1	0.89	0.32	0.48 – 1.65	0.704	1.17
Access (vs Retention) Cost x RB	2	1.40	0.26	0.83 – 2.35	0.205	0.75
Binding (Fixed vs Random) Cost x RB	1	1.60	0.22	1.05 – 2.44	**0.029**	0.09
Control vs Rest	*2*	*0.40*	*0.20*	*0.27 – 0.59*	***<0.001***	*0.21*
Access vs Rest	*1*	*0.26*	*0.17*	*0.19 – 0.36*	***<0.001***	*0.30*
Control vs Rest x RB	*2*	*1.11*	*0.22*	*0.72 – 1.71*	*0.64*	*0.31*
Access vs Rest x RB	*1*	*1.32*	*0.18*	*0.93 – 1.88*	*0.13*	*0.15*
N = 1,496 observations; Conditional R^2 = .375						**σ^2= 3.29**

Notes: To test the contrast of interest, two sets of orthogonal contrasts were needed. The column Model indicates which model the estimates have come from. Binding was in both models and as expected, produced identical estimates for all effects in both models.

Model 1: glmer(ACT ~ 1 + bindingC*RIcomposite2 + retentionC*RIcomposite2 + AccessVrestC*RIcomposite2 + (1 + bindingC*RIcomposite2 + retentionC*RIcomposite2 + AccessVrestC*RIcomposite2 | subject)

Model 2: glmer(ACT ~ 1 + bindingC*RIcomposite2 + accessC*RIcomposite2 + ControlVrestC*RIcomposite2 + (1 + bindingC*RIcomposite2 + accessC*RIcomposite2 + ControlVrestC*RIcomposite2 | subject)

Table B2. *Effect contrast coding for Models 1 and 2*

Model 1 Effect Contrast Coding	**C**	**R**	**AF**	**AR**
bindingC	0	0	-0.5	0.5
retentionC	-0.5	0.5	0	0
AccessVrestC	-0.5	-0.5	0.5	0.5
Model 2 Effect Contrast Coding	**C**	**R**	**AF**	**AR**
bindingC	0	0	-0.5	0.5
accessC	0	-2/3	1/3	1/3
ControlVrestC	-3/4	1/4	1/4	1/4

RIcomposite2 = standardized RB

APPENDIX C

For the experiment presented in Chapter VII on the Change Detection task, the original task extended from the base 12 ‘practice’ items into another ‘test’ set of a further 32 items. As discussed in Chapter VIII, this test data was contaminated by an overly powerful “ignore instruction” manipulation. The ignore instruction appeared either before or after exposure to the probe display and instructed participants to ‘ignore’ any changes that occurred to a certain type of shape (e.g., “triangle”).

The data presented here are the results of analyses on the test set, documented here in the interest of transparency.

The overall proportion of correct trials for the Test set was .799 for same items, .598 for cluster changes, and .629 for single-object changes. Overall, difference in accuracy for single-objects changes compared to cluster changes was significant on a paired-sample *t*-test, t_{951} = 3.68, $p < .001$, but the effect was small, $d = .15$ (in contrast to $d = .32$ for the Practice set). In comparing the two sets in Figure C1, it is clear that that the addition of ignore instructions meant participants struggled more with the Test Set than the Practice Set, and Figure C2 indicates this is specifically due to the post-probe instructions. Using a repeated-measures ANOVA, the interaction between target change type (cluster vs. single-object) and set (practice vs. test) was indeed significant, $F_{1,951}$ = 20.90, $p < .001$, η_p2 = .022. Although the interaction itself is consistent with H2 (in that cluster and single change performance become more similar), the direction of the effect indicates that performance on single changes dropped to match cluster performance (rather than cluster improving to match single). However, as seen in Figure C2, this was (again) largely a result of the impact of the timing of the ignore instruction. In general, we underestimated the impact of this timing variable. When considering only *pre-probe* Test items, the H2 effect was in the hypothesized direction, with cluster changes (M = .696) improving to match single changes (M = .700). That is, pre-probe Test single changes (M = .700) were more similar to Practice single changes (M =.729) than pre-probe Test cluster changes (M = .696) were to Practice cluster changes (M = .636), $F_{1,951}$ = 32.92, *p* < .001, $\eta_p 2$ = .033. In other words, H2 did appear to come to fruition in the test set, but only when considering pre-probe.

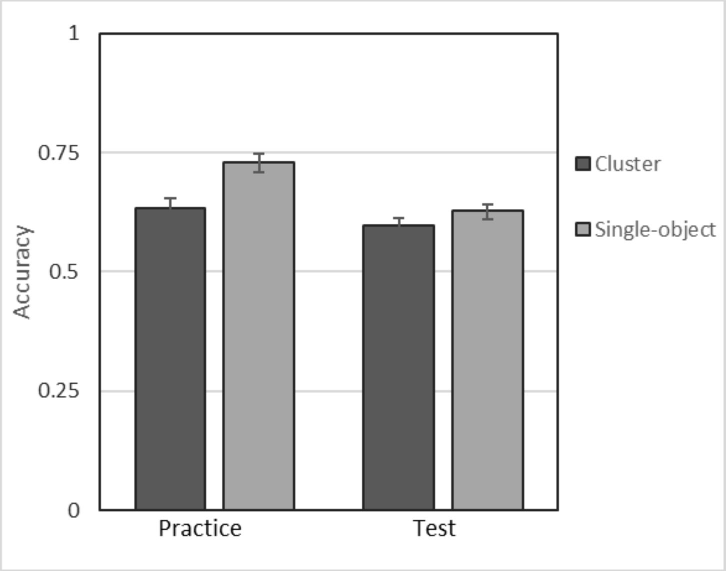


Figure C1. Average proportion correct for target changes types in Practice Set compared to the Test Set. Error bars represent 2 x standard errors.

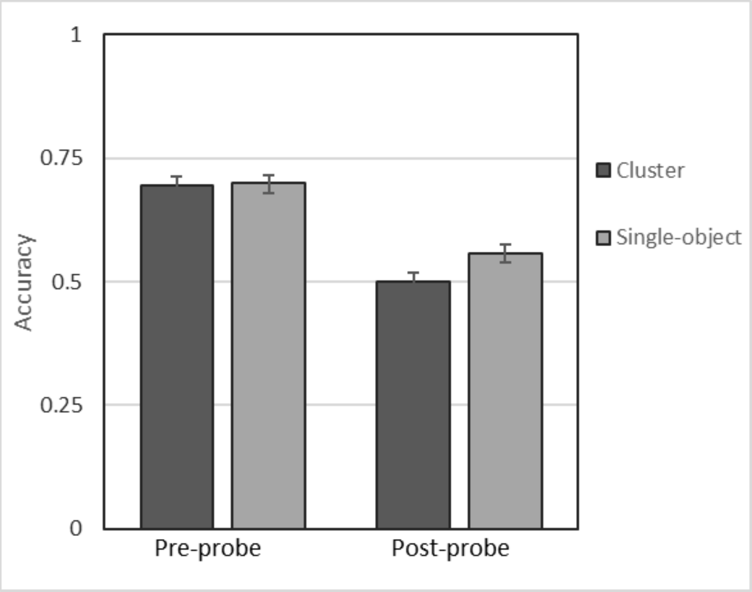


Figure C2. Average proportion correct for target changes types in the Test Set, broken down by (a) direction and (b) exposure. Error bars represent 2 x standard errors.

As seen in Figure C3, the pattern of results for H3 (Figure C3b) and H4 (Figure C3a) largely followed the (non-significant) pattern of results from the Practice items. For H3, a mixed ANOVA was conducted (with change type as a within-subjects variable and exposure as a between-subjects variable). The difference between single and cluster changes was no different for 5 seconds compared to 3 seconds (Figure C3b), $F_{1,948} = 2.06$, $p = .152$, consistent with the Practice items (i.e., both Practice and Test analyses are contrary to H3). The catastrophic impact of pre-probe vs. post-probe ignore instructions (see Figure C2) meant that it was necessary to clarify the H3 and H4 Test set effects by adding instruction timing as a within-subjects variable. However, this three-way interaction (exposure x change x timing) was also not significant, $F_{1,948} = 1.86$, $p = .173$, indicating that – regardless of timing – exposure did not influence the difference between cluster and single performance.

For H4, an independent samples *t*-test was conducted. The difference between distancing and uniting participants on single change performance was significant, $t_{950} = 2.43$, $p = .015$, but the effect was small, $d = .16$. This small difference may indicate that the results were following the three-way interaction discovered by the Practice set which included trial order, where the difference between distancing and uniting for single changes was reducing as the task went on. However, when adding timing as a within-subjects variable, the mixed ANOVA revealed a two-way interaction between timing and direction on single change performance, $F_{1,950} = 34.06$, $p < .001$, $\eta_p2 = .035$, such that the difference between distancing and uniting was larger for pre-probe (distancing M = .742; uniting M = .648) than for post-probe (distancing M = .545, uniting M = .574). Thus, single change performance was indeed significantly worse for uniting than distancing, but only for pre-probe. Post-probe, meanwhile, appeared to be suffering a floor effect.

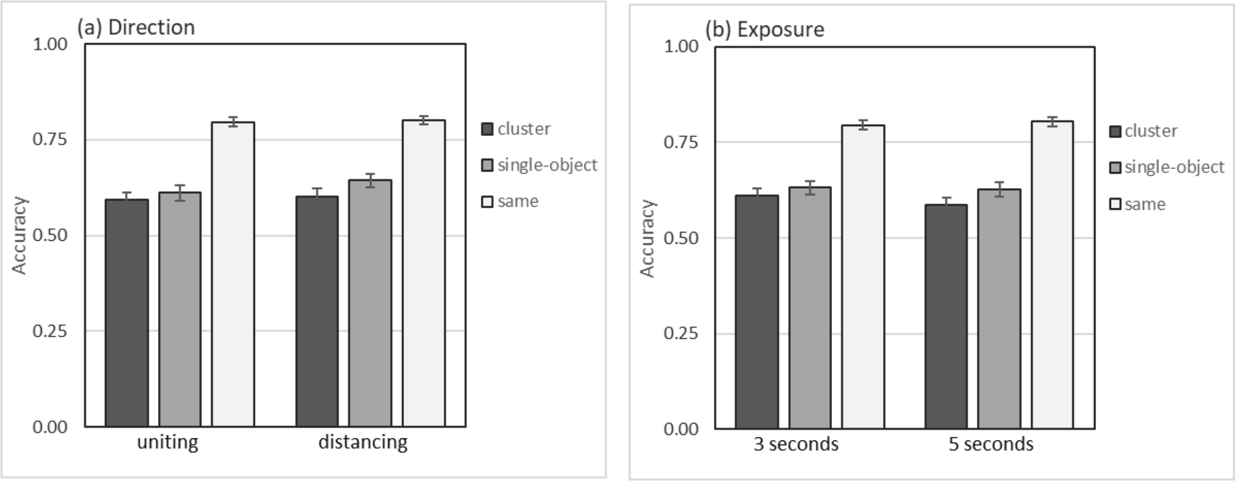


Figure C3. Average proportion correct for target changes types in the Test Set, broken down by (a) direction and (b) exposure. Error bars represent 2 x standard errors.

As described in the prior section, the impact of pre-probe vs. post-probe on the between-subjects variables was clear. Indeed, the difference in accuracy between pre-probe (M = .767, SD = .125) and post-probe (M = .645, SD = .129) was significant, $F_{1,948}$ = 521.34, *p* < .001 (consistent with H5), and disruptively powerful, $\eta_{p}2$ = .355. The interaction between target change and instruction timing was also significant, $F_{1,948}$ = 26.33, *p* < .001, $\eta_{p}2$ = .027, such that the difference between cluster and single-object changes was larger for post-probe trials than for pre-probe trials. In fact, as seen in Figure 7.8, pre-probe trials had virtually equivalent performance between the two types of change; while post-probe instructions caused cluster changes to fall to chance level (50%). This outcome thus seems to suggest that the difference in performance between the two types of target changes (cluster and single-object) did indeed minimize (even equivalize) after the Practice Set after all (consistent with H2); but post-probe instructions were so difficult that the single-object advantage resurfaces (but only slightly rising above the chance-level performance seen in cluster changes).

The catastrophic effect of timing would clearly impact on H6 (synchronous ignore~target changes should be superior for post-probe compared to pre-probe; desynchronous changes should be superior for pre-probe to post-probe) because pre-probe would always have superior performance to post-probe regardless of other variables.

However, it would still be of interest to see if the gap between pre- and post-probe is smaller in synchronous items compared to desynchronous items, because we would expect the detrimental impact of post-probe is weakened if the ignored objects (which are harder to unbind than they are to proactively ignore) are in sync with the target objects. A repeated measures ANOVA was conducted and, although the timing x synchronicity interaction was significant, $F_{1,951} = 128.34$, $p < .001$, $\eta_p2 = .119$; it was in the opposite direction to the prediction of H6. That is, the gap between pre-probe and post-probe accuracy was larger for in-sync changes (pre-probe M = .770, post-probe M = .590) compared to out-of-sync changes (pre-probe M = .764, post-probe M = .697). However, note that the purpose of this synchronicity analysis was solely the interaction with timing. Although it appears that synchronicity results in a general disadvantage (in-sync M = .680, out-of-sync M = .729), this result cannot be taken in isolation because in-sync items are weighted more heavily towards ‘different’ items which are more difficult than ‘same’ items. When adjusting for this bias by weighting the same items equally to the different items, the expected advantage of synchronicity emerges (in-sync M = .740, out-of-sync M = .681); and the interaction with timing remains the same, with a larger gap between pre- and post-probe for in-sync, compared to out-of-sync.

H7 and H8 were comparatively simple analyses, since they focus solely on false alarm rates for ‘same’ items (i.e., participant saying ‘different’ to a ‘same’ item). A repeated-measures ANOVA indicated that changes to the ignored objects did change the false alarm rates for ‘same’ items, $F_{3,2853} = 66.36$, $p < .001$, $\eta_p2 = .065$. Two planned contrasts were conducted with Bonferroni correcting the error rate to .025. The first contrast comparing cluster to single-object changes was significant, $F_{1,951} = 8.55$, $p = .004$, but weak, $\eta_p2 = .009$, indicating a slightly higher false alarm rate for single changes compared to cluster changes (consistent with H7). Consistent with H8, the second contrast indicated that scatter changes resulted in significantly more false alarms than the other three conditions on average, $F_{1,951}$ = 114.81, p < .001, and this effect was powerful, η_p2 = .108. Figure C4 depicts these false alarm rates, demonstrating a higher false alarm rate for scatter objects.

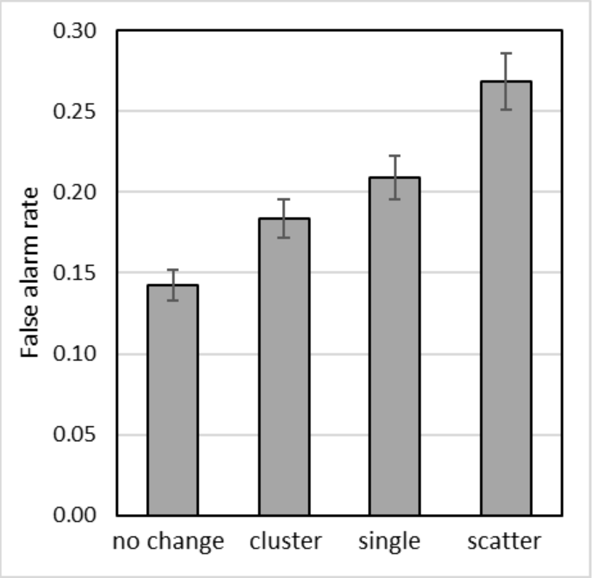


Figure C4. False alarm rates for ‘same’ items based on the change in ignored objects. Error bars represent 2 x standard error.