Chapter 4: Meta-analyses of Errors

Empirical Constraints on Error Distributions

This chapter describes three meta-analyses of a number of recent, unpublished studies on short-term serial recall conducted at the Applied Psychology Unit. These analyses were to test the generality of the results of Experiments 1-3, and also to test more detailed predictions of three specific models of serial recall, a chaining model (Murdock, 1995), a positional model (Burgess & Hitch, 1992) and an ordinal model (Page & Norris, 1996b). Meta-analyses over several different experiments were necessary because some types of errors are rare (e.g., repetitions and protrusions), hampering statistical tests within any one experiment. The results are summarised in a set of *empirical constraints* on error distributions in serial recall.

Three Models of Serial Recall

In order to guide some aspects of the meta-analyses, three specific models of serial recall were considered, each of which exemplified one of the general theories of serial order in Chapter 1. The first was a closed-loop, compound chaining model, based on the Power Set Model of Murdock (1995). Though Murdock did not specify the precise nature of the closed-loop chaining (i.e., whether or not errors are fed back as cues), such a closed-loop model has been analysed independently by Henson (1994). The second model was a positional model based on the Articulatory Loop Model of Burgess and Hitch (1992). This model uses a context signal to cue each position, such that cues for nearby positions overlap in symmetrical manner (Chapter 1; decay processes in this model were ignored for simplicity). The third model was an ordinal model based on the Primacy Model of Page and Norris (1996b). This assumes a primacy gradient of activations, invariant across positions (decay was again ignored).

Competition Space

Though the above models differ in many respects, they can be compared using the abstract notion of a *competition space*. Competition space indicates the strength with which each item competes for each response during serial recall. The competition space for the first three responses in serial recall of five items is shown in Figure 4-1. The filled bars represent the strength with which each item (from left to right) competes for the first, second and third response (in each column). Assuming some random noise in these strengths, the height of each bar relative to the others relates to the probability of recalling that item at that position (i.e., the random noise can sometimes cause errors).

The broken bars represent items that have been recalled. Thus Figure 4-1 illustrates competition space at the start of recall (leftmost column), after Item 2 has been recalled erroneously in Position 1 (middle column), and after Item 3 has been recalled erroneously in Position 2 (rightmost column). Once recalled, an item is suppressed. This suppression reduces the probability of recalling it again, explaining why repetitions are rare (below). Suppression also explains the interdependency between responses (Henson et al., 1996), such that the probability of recalling an item depends on what has been recalled previously (i.e., items are selected without replacement). All three models above assume a process of suppression (implicitly in Murdock, 1995; explicitly in Burgess & Hitch, 1992, and Page & Norris, 1995).

In competition for the first response (leftmost column), it is assumed all three models are equivalent.[1] In other words, all models predict the first item as the most likely response, the second item as the next most likely, etc. The first difference between the three models arises when the first error occurs, where Item 2 is recalled in Position 1 (middle column). The models differ in their predictions as to what should follow this error. The Power Set Model predicts the most likely next response is Item 3, because it will be cued strongly by previous recall of its associate, Item 2. The Articulatory Loop Model predicts that Item 1 and Item 3 will be equally likely to follow, because the cue for Position 2 overlaps equally with those for Position 1 and Position 3 (e.g., Figure 1-2 in Chapter 1). The Primacy Model predicts that Item 1 will be most likely to follow, because it remains the strongest competitor.

The prediction of the Primacy Model, that Item 1 will follow the error on Position 1, was termed *fill-in* by Page and Norris (1996b). In more general terms, fill-in is a property such that "when an item is missed out in recall, due to a transposition, it is liable to be recalled in the next position" (Page & Norris, 1996b, p. 8). Fill-in is important in preventing a cascade of further errors. This is evident by considering the situation where Item 2 is followed by a

1. An additional start-of-list context is assumed in order to cue the first item in the Power Set Model.

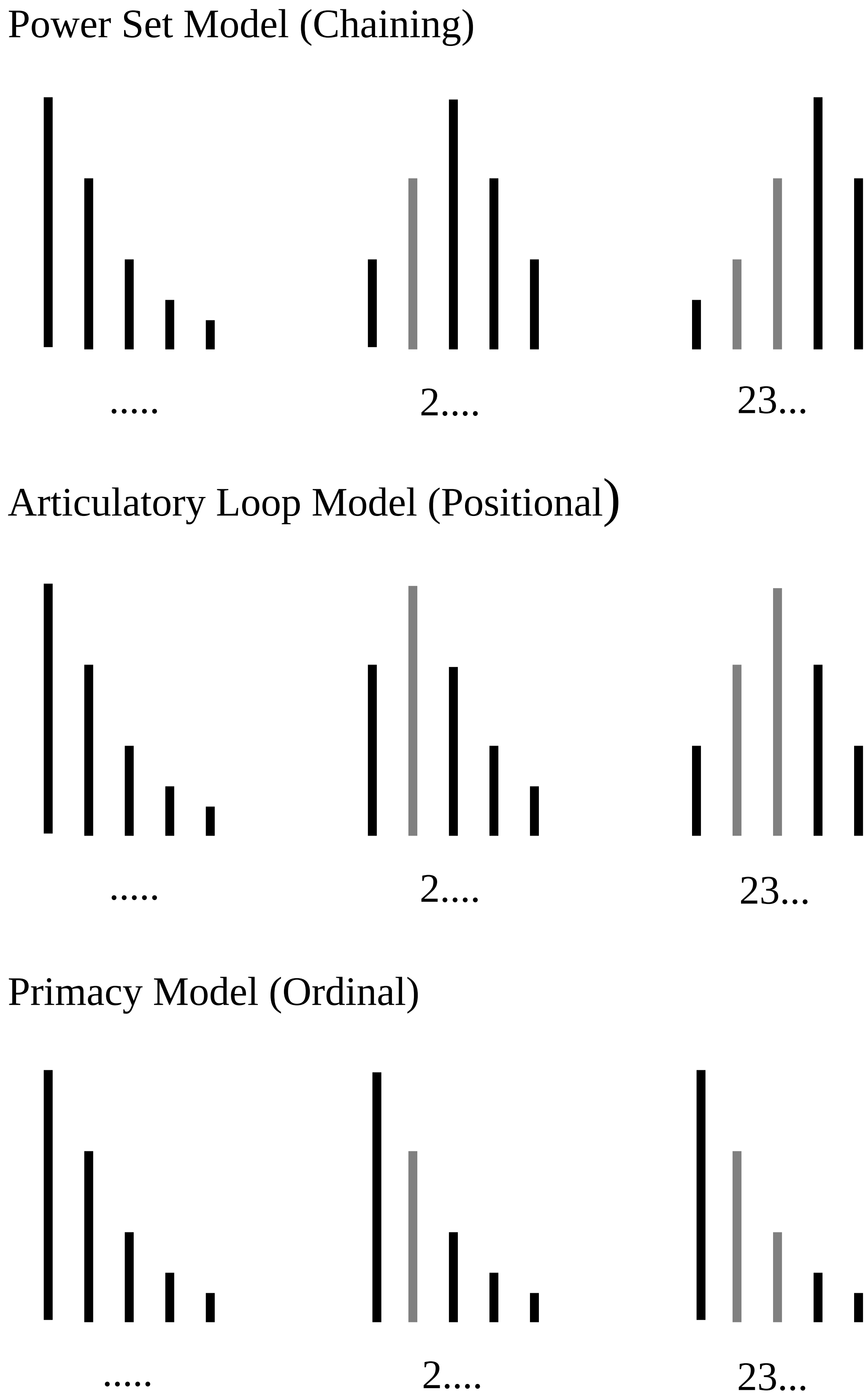

Figure 4-1: Competition space within each model for the first three responses to a list *12345* recalled as *23...*, illustrating absence of weak fill-in.

further error of Item 3 (rightmost column). The Power Set Model predicts the most likely next response is Item 4, because it will be cued strongly by previous recall of its associates Item 2 and Item 3. In other words, the Power Set Model predicts a further slippage of items. The Articulatory Loop Model also predicts Item 4 will follow, because the cue for Position 3 overlaps more with the cue for Position 4 than the cue for Position 1. Only the Primacy Model predicts that the most likely next response remains Item 1, to “fill-in” the gap and prevent further slippage. In other words, only the Primacy Model predicts that the probability of fill-in increases with further errors; the other models predict that the probability of fill-in decreases, such that the last item is unlikely to be recalled until the end, when all others have been recalled and suppressed. Further consideration reveals that the lack of fill-in in the Power Set and Articulatory Loop Models is why neither produce sufficient recency (Henson et al., 1996).

On the other hand, consider the situation in Figure 4-2, where the first two items have been recalled correctly (leftmost column). The middle column then shows the competition space after Item 5 is recalled erroneously in Position 3. The Power Set and Articulatory Loop Models predict the next most likely response is the correct response, Item 4, whereas the Primacy Model predicts fill-in of Item 3.

This example illustrates the distinction between strong fill-in and weak fill-in. The Primacy Model shows strong fill-in, in that the earliest unrecalled item will always be the most likely response following an error. The Power Set and Articulatory Loop Models do not predict strong fill-in. The Articulatory Loop Model in particular predicts that the correct response is always most likely following an error (providing the correct item has not already been recalled and suppressed, as in Figure 4-1). Only in situations where the correct item has already been recalled can (weak) fill-in can occur, as in the rightmost column of Figure 4-2. In this case, Item 4 is recalled correctly in the fourth position, and all three models predict that Item 3 will finally fill-in.

The distinction between strong and weak fill-in is important because a model can show weak fill-in without showing strong fill-in. Though this is not true of the Articulatory Loop Model, because of the symmetrical nature of its positional cue (a situation actually made worse once decay is added, Henson et al., 1996), it is true of the new positional model

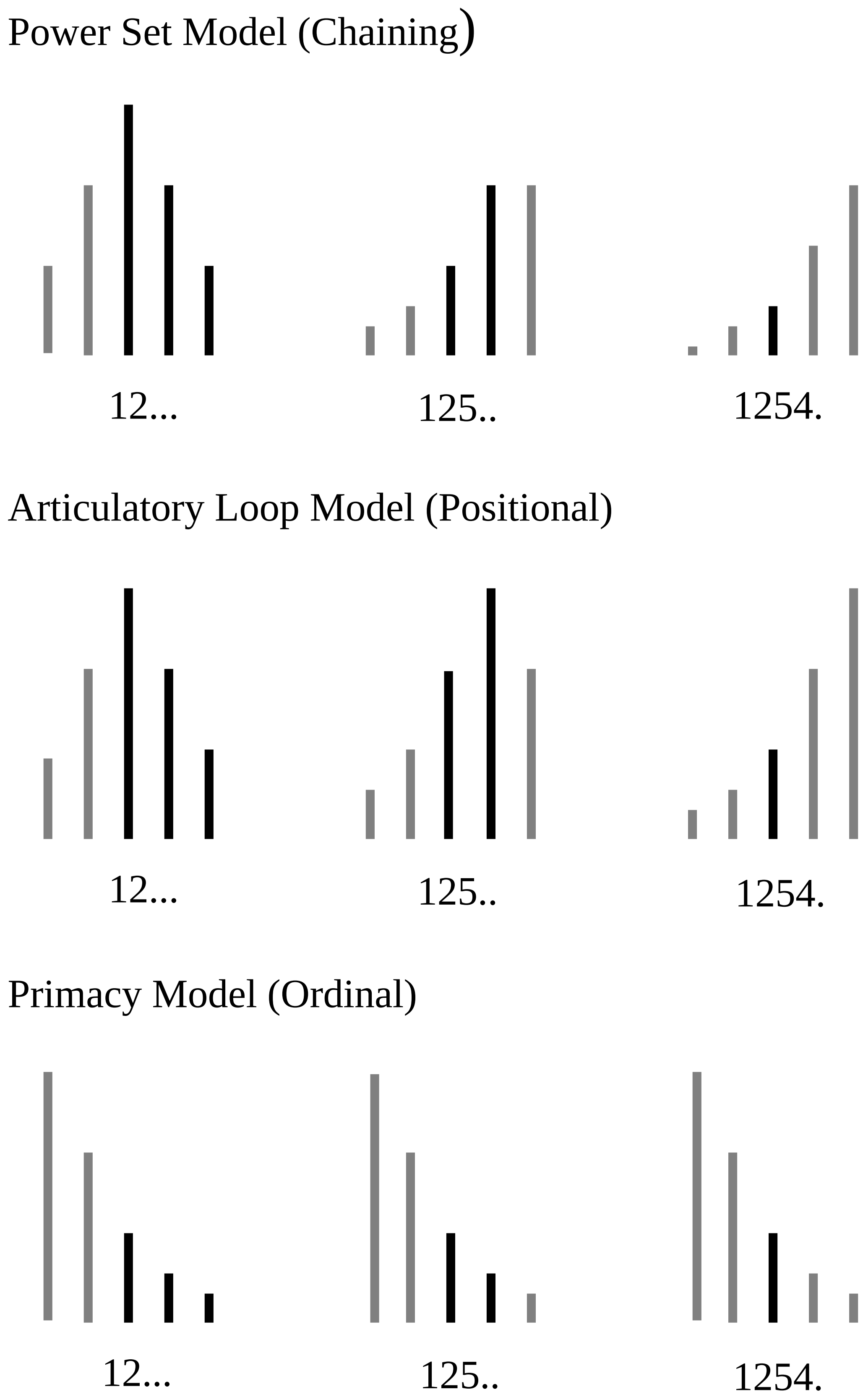

Figure 4-2: Competition space within each model for the last three responses to a list *12345* recalled as *1254.,* illustrating absence of strong fill-in.

developed in Chapter 5. This model assumes asymmetrical positional cues, biased towards earlier items. The strength of fill-in was one of the questions asked in the meta-analysis below.

Meta-analysis 1

In total, 37 conditions from 14 different experiments were analysed, using a computer program developed by the author. These experiments all employed immediate serial recall of objectively ungrouped lists of phonologically dissimilar, nonrepeated items. The conditions differed in list length (from five to nine items), nature of items (digits, letters or words), presentation rates (between one and two items per second), presentation modality (visual, vocalised, or auditory) and recall method (written or spoken). Further details of the conditions are given in Appendix 2. A number of pairwise, binomial sign-tests were performed across the conditions, accompanied by 95% confidence intervals (*CI*) for the median value.

Primacy Constraint

Primacy was evident in all error position curves in Experiments 1-3. Its prevalence was tested further by comparing the frequency of errors on the first two positions across all 37 conditions. Errors on the first position were less frequent than on the second position in all cases, $N=37, p<.001, CI=(.11,.15)$, reinforcing the ubiquity of primacy in serial recall.

Recency Constraint

Recency was also evident in Experiments 1-3, though it was weaker than primacy, and confined to the last one or two positions. Last-item recency was tested by comparing the frequency of errors on the last two positions. The frequency on the last position was less than on the penultimate position in only 20 conditions (and equal in 4 conditions), suggesting that recency is not a reliable effect in serial recall, $N=33, p=.15, CI=(-.02,.04)$.

However, of the 13 conditions with no last-item recency, 12 employed lists of words, and 7 of these used five-syllable words. These conditions showed large increases in omissions towards the end of recall (Experiment 2; below). When the 10 conditions with words of more than one syllable were excluded from analysis, the presence of last-item recency was reliable, arising in 19 conditions (and equal in 2 conditions), $N=25, p<.01, CI=(.01,.10)$. This suggests that recency is normally found, except when there are large numbers of omissions.

Locality Constraint

The locality constraint, that items transpose small distances about their correct positions, was introduced in Experiment 1. The generality of this constraint was tested by comparing the frequency of one-and two-apart transpositions, weighted by the opportunity for such transpositions.[2] One-aparts were more frequent than two-aparts in all conditions, $N=37$, $p<.001$, $CI=(.03,.04)$, demonstrating the fundamental nature of the locality constraint.

Fill-in Constraint

It is possible for data (and models) to meet the locality constraint without meeting the fill-in constraint (above). For example, a sequence *12345* recalled as *13452* contains three one-apart transpositions, and one three-apart transposition. Though the ratio of these transpositions would meet the locality constraint, this example violates the fill-in constraint because Item 2 was not recalled immediately after it was replaced by Item 3.

To measure fill-in, analysis was confined to responses following the first error in a report. To illustrate the nature of such responses, data from the ungrouped conditions of Experiment 2 were collapsed across subjects (Table 4-1). Of the 207 responses following a first error of Item $i+1$ on Position i (as in Figure 4-1), the majority were the fill-in errors of Item i predicted by the Primacy Model, and only half as many were the associate errors of Item $i+2$ predicted by the Power Set Model (top row of Table 4-1). In other words, when $i=1$, an incorrect report of *12345* is more likely to be *21345* than *23145* (contrary to Figure 4-1). Thus, there was evidence for fill-in. There were hardly any immediate repetitions of the correct Item $i+1$, but this is attributable to the suppression of items already recalled.

	Following Response			
First Error	Fill-in (Item i)	Correct (Item $i+1$)	Associate (Item $i+2$)	Other
Item $i+1$	.53	.01	.21	.25
Item $j>i+1$	.25	.48	.08	.19

Table 4-1: Proportion of responses following a first error on Position i in Experiment 2.

2. Given transposition distances of i and j $(i<j)$, this weighting means scaling the number of j-apart transpositions by a factor $(n-i)/(n-j)$, where n is the list length, reflecting the fewer opportunities for transpositions further apart.

To measure the strength of fill-in, analysis was confined to responses following a first error of Item *j* on Position *i*, where Item *j* was an item other than Item *i* or Item *i+1* (as in Figure 4-2). Of the 336 such responses, the majority were the correct responses of Item *i+1* predicted by the Articulatory Loop Model, and only half as many were the fill-in errors predicted by the Primacy Model (bottom row of Table 4-1). In other words, when *i=3*, incorrect report of *12345* is more likely to be *12543* than *12534* (as in Figure 4-2). Thus, there was no evidence for strong fill-in.

To test the generality of this conclusion, similar calculations were performed in the meta-analysis. With a first error of Item *i+1*, the proportion of following responses that were fill-in errors was greater than the proportion that were associate errors in 35 conditions (equal in 2 conditions), demonstrating highly reliable weak fill-in, $N=35$, $p<.001$, $CI=(.20,.32)$. With a first error of Item *j*, where $j>i+1$, the proportion of following responses that were correct was greater than the proportion that were fill-in errors in 33 conditions (equal in 1 condition), $N=36$, $p<.001$, $CI=(.16,.22)$, demonstrating that strong fill-in is the exception rather than the rule. Taken together, these analyses confirm that the fill-in is stronger than predicted by the Articulatory Loop Model, but not as strong as predicted by the Primacy Model.

One caveat applies to the above analysis. Many of the lists are likely to be spontaneously grouped (Chapter 3). The influence of grouping may confound the analysis, perhaps reducing the strength of fill-in, given that interpositions tend to be followed by correct responses (Experiment 2). The fact remains however that the interpositions themselves, or indeed any type of positional error, cannot be explained by models with strong fill-in (below).

Omission Constraint

The omissions in Experiments 2 and 3 increased towards the end of recall. To test the reliability of this finding, the frequency of omissions on the last two output positions was compared. The frequency was greater on the last position than penultimate position in 32 conditions (equal in 3), confirming the reliability of the finding, $N=34$, $p<.001$, $CI=(.04,.17)$.

The fact that omissions increase towards the end of recall might suggest that the last item is omitted more often than any other. Indeed, this is what is predicted by the Primacy Model (Page & Norris, 1996b). To illustrate this, the upper panel of Figure 4-3 shows the

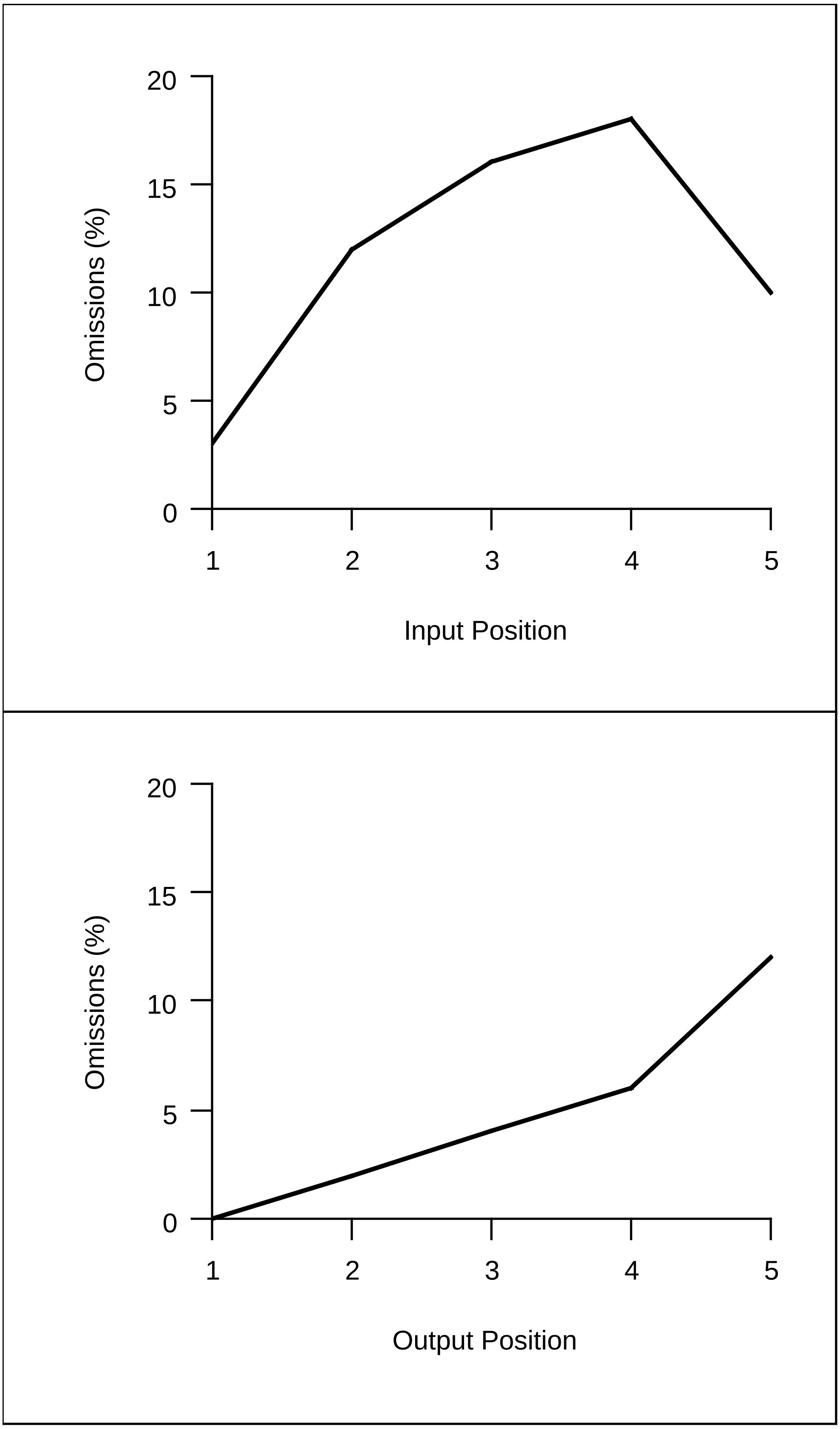

Figure 4-3: Omissions by input position (upper panel) and by output position (lower panel) averaged across both conditions of Experiment 3.

frequency of omissions against input position averaged over both conditions of Experiment 3. The increase in omissions with output position (lower panel) was not paralleled by a similar increase with input position: The last item was more often recalled somewhere than the penultimate item. This pattern of results can be explained if the last item is sometimes recalled too early, replacing the penultimate item, and followed by an omission. To test whether the pattern was an exception rather than the rule, the meta-analysis compared the frequency of omissions on the last input position with the frequency of omissions on the penultimate input position. The frequency on the last position was greater than on the penultimate position in 17 conditions (and equal in 3 conditions), *N=34*, *p=.5*7, *CI=(-.02,.03)*. This unreliable difference indicates that the increase in omission towards the end of recall does not always reflect failure to recall the last item. As well as being troublesome for the Primacy Model, this pattern of item errors contrasts with the flat distribution assumed by Lee and Estes (1977, 1981). This is probably because they, like Healy (1974), did not consider lists of more than four items.

Repetition Constraint

Repetitions in Experiments 1-3 were rare. However, their distribution was highly constrained: They were always widely separated in reports, with the majority being items recalled at the start of recall that were recalled again towards the end of recall. In condition PN of Experiment 1 for example, repetitions comprised approximately 2% of responses (11% of errors) and the two occurrences were, on average, 3.34 positions apart in reports. The most common repetition was of the first item, recalled correctly on Position 1 and again incorrectly on Position 6 (hence the exception to the locality constraint for Position 6 of this condition in Experiment 1). This pattern is shown in Figure 4-4 (the peak on the fourth input position probably reflects the effect of the 3-3, subjective grouping in Experiment 1).

The significance of this distribution of repetitions can be illustrated by a simple guessing model. According to this model, subjects who fail to recall an item correctly guess at random from the set of list items. Simulations of such a simple model, fitted to overall error rates in condition PN, produced repetitions that comprised 16% of responses (84% of errors), far in excess of the data. Simulations also gave a mean distance between two occurrences of an item of 2.21 positions, considerably smaller than in the data. Though a different frequency of

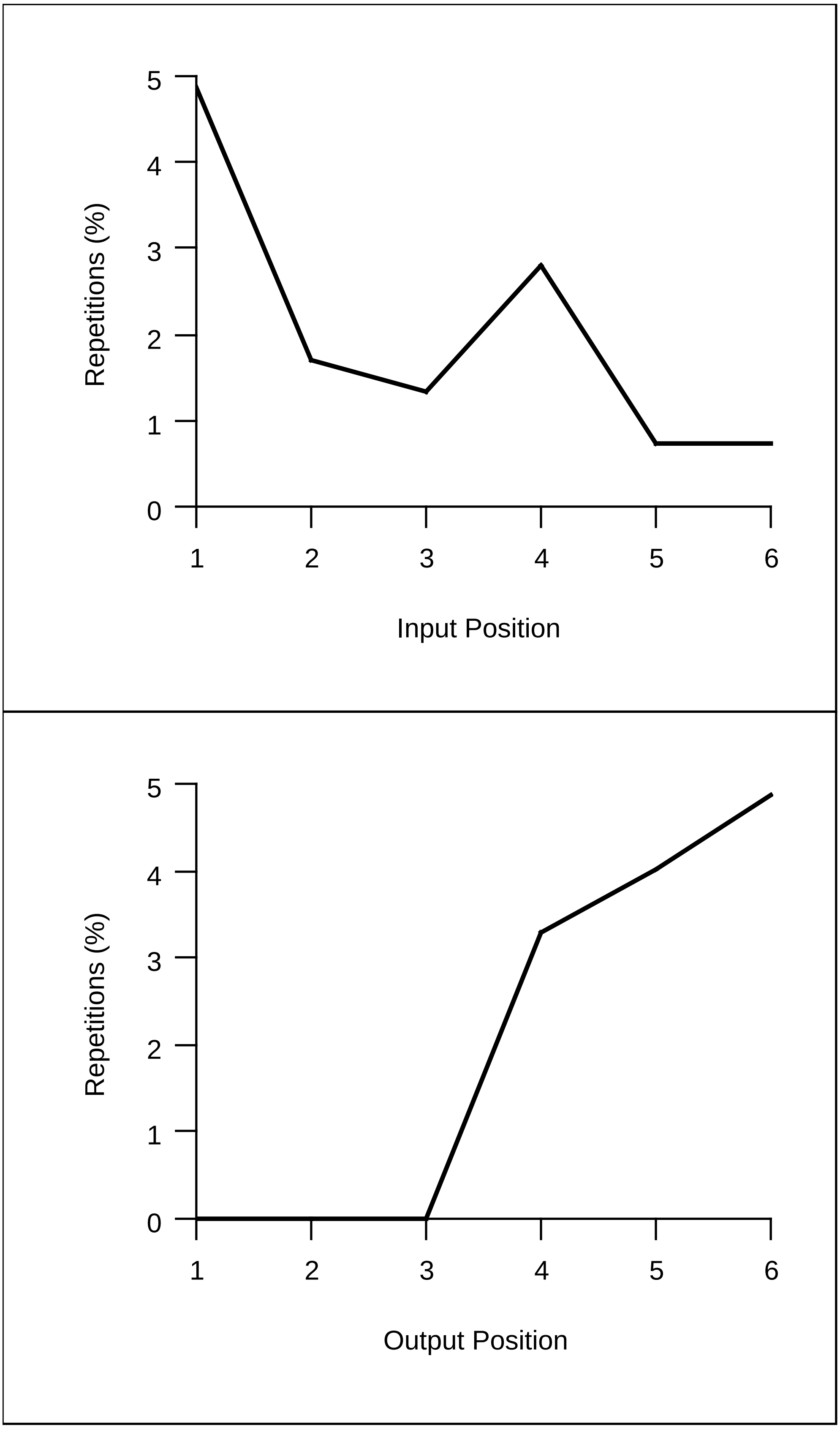

Figure 4-4: Repetitions by input position (upper panel) and by output position (lower panel) for PN condition of Experiment 1.

repetitions would result if subjects' guesses were biased towards neighbouring list items, as required by the locality constraint, this would produce an even smaller mean distance between the two occurrences of a repeated item. This example suggests that responses in serial recall are normally chosen without replacement, supporting the idea of response suppression described above. Nonetheless, the fact that repetitions do sometimes occur suggests that response suppression is not perfect; it probably wears off over time (Chapter 5).

The hypothesis that most repetitions involve early list items was examined by testing whether over 50% of repetitions were from the first two input positions. This was true of 24 conditions in the meta-analysis, confirming the hypothesis, *N=37, p<.05, CI=(.01,.19)*. The hypothesis that most repetitions occur towards the end of recall was examined by testing whether over 50% of repetitions occurred on the last two output positions. This was true of 26 conditions (equal in 1 condition), confirming this hypothesis too, *N=36, p<.01, CI=(.02,.17)*.

Protrusion Constraint

The protrusions measured in Experiment 3 were also rare. Nonetheless, they represented a significant proportion of immediate intrusions; a proportion greater than expected by chance. To test whether this was true more generally, the proportion of erroneous items that occurred at the same position in the previous report, given that they occurred somewhere in that report, was compared with that expected by chance (which is *1/n*, where *n* is the list length). This proportion was above chance in 35 conditions (and equal in 1 condition), demonstrating that output protrusions are a reliable finding, *N=36, p<.001, CI=(.06,.10)*. Moreover, this proportion was greater than the corresponding proportion for input protrusions in 28 conditions (and equal in 3 conditions), *N=34, p<.001, CI=(.01,.03)* supporting the suggestion in Experiment 3 that output protrusions are a better index of positional information. Finally, the proportion of output protrusions followed by a correct response was greater than the proportion followed by a further protrusion in 35 conditions, *N=37, p<.0001, CI=(.11,.20)*, supporting the conclusion of Experiment 3 that protrusions normally arise singly, without intrusion of whole subsequences.

The theoretical importance of positional errors like protrusions can also be illustrated in competition space. Figure 4-5 shows the competition space in recall of the second of five

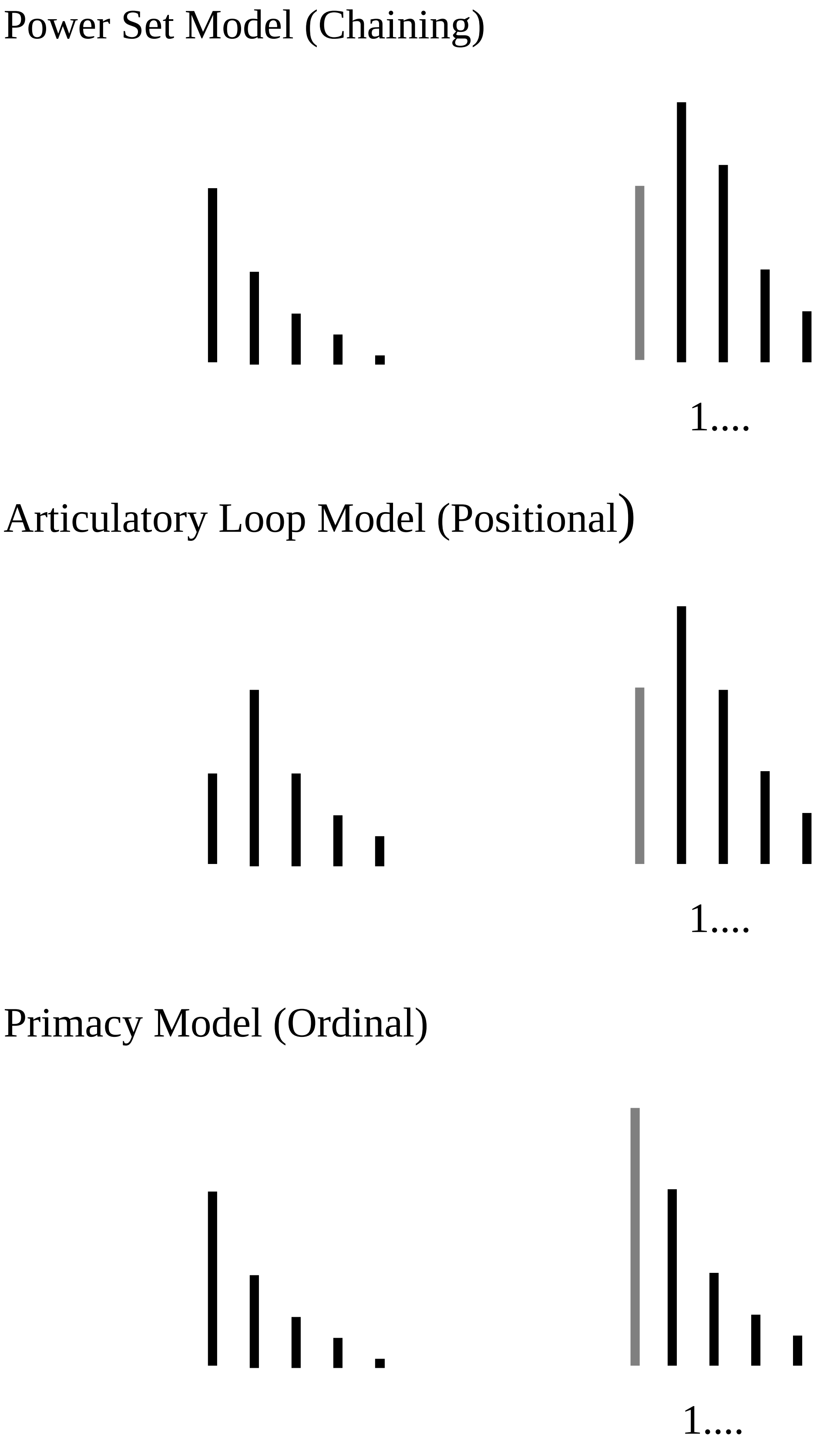

Figure 4-5: Competition space within each model for the second response to a list *12345* recalled as *1....*, illustrating competition from items in the previous trial.

items (rightmost column), including competition from items in the previous trial (leftmost column), assuming suppression for the previous trial has worn off. In the Power Set Model, the cue for the second item bears no necessary resemblence to the cue for the second item of the previous trial (unless the first item happened to be the same in both trials). Thus, if there is to be an intrusion from the previous trial, there is no reason for it to be an protrusion of the second item from that trial. (In Figure 4-5, the most likely intrusion is the first item of the previous trial, assuming that the two lists share remote associations with the same start-of-list context). A similar argument applies to ordinal models like the Primacy Model, because a start-of-list cue (Page & Norris, 1996b) would mean that the most likely intrusion is always the first item from the previous trial. Only the Articulatory Loop Model predicts that the most likely intrusion is a protrusion, as in the data. This is because only a positional model assumes separate cues for each position, and, assuming the same cues are reused on each trial, any proactive interference will be of a positional kind. This illustrates the point made in Chapters 1 and 3, that positional errors necessitate a positional theory.

Meta-analysis 2

This meta-analysis examined 9 conditions from 9 different experiments with grouped lists of phonologically dissimilar, nonrepeated items, to test the reliability of the results of grouping in Experiment 2. Further details of the conditions are given in Appendix 2.

Interposition Constraint

The grouped condition of Experiment 2 showed a greater proportion of three-apart interpositions than two-apart transpositions. To test the reliability of this finding, the frequency of transpositions n positions apart (with groups of size n) was compared to the frequency of transpositions $n-1$ positions apart, weighted by the opportunity for such transpositions (Footnote 2). The proportion of interpositions was greater in all 9 conditions, $N=9$, $p<.005$, $CI=(.01,.02)$. This confirms that interpositions in grouped lists override the locality constraint.

The grouped condition of Experiment 2 also demonstrated that more interpositions arose between the middle of groups than the start or end of groups. This finding was confirmed by comparing the proportion of interpositions between the middle of groups with the proportion between the start and end of groups. The proportion on middle positions was greater in all conditions, *N=9, p<.005, CI=(.02,.07)*. Finally, it was also confirmed that the proportion of interpositions followed by a correct response was greater than the proportion followed by a further interposition in all conditions, *N=9, p<.005, CI=(.28,.45)*, supporting the conclusion that interpositions, like protrusions, arise singly.

Meta-analysis 3

This meta-analysis examined 10 conditions from 3 different experiments that employed ungrouped lists in which phonologically similar and phonologically dissimilar items alternated. This was to test the reliability of the findings of Experiment 1. Further details of the conditions are given in Appendix 2.

Confusion Constraint

Experiment 1 demonstrated that phonologically confusable items tend to transpose with one another, causing more errors for confusable items than nonconfusable items in lists where they alternate. All 9 conditions in the meta-analysis also showed a higher frequency of errors for confusable than nonconfusable items, *N=10, p<.005, CI=(.12,.22)*. However, Experiment 1 failed to find a consistent effect of confusable items on the recall of alternated nonconfusable items. This failure prompted two conclusions: 1) there is no effect of phonological similarity on cuing, and 2) there is no effect of errors on cuing (Chapter 2).

To test this finding, the frequency of errors on nonconfusable positions in alternating curves was compared with that in nonconfusable curves. There was a higher frequency of errors on nonconfusable positions in alternating curves in 8 of the 10 conditions, a result that was almost reliable, *N=10, p=.05, CI=(.02,.07)*. This suggests the first finding in Experiment 1 may not generalise, particularly for lower-span subjects. One possible reason for this is the general knock-on effects of errors (Experiment 1). To test this notion, the above errors were conditionalised on correct recall of preceding items (Henson et al., 1996). In this case, the conditional probability of errors on nonconfusable positions in alternating curves was greater than in nonconfusable curves in only 4 conditions (and equal in 1 condition); a result that was not reliable, *N=9, p=.75, CI=(-.01,.01)*. This is consistent with the knock-on effects of an error (though also consistent with an effect of errors on cuing). More importantly, it is inconsistent with an effect of similarity on cuing, ruling out most chaining models (Chapter 1).

Finally, Experiment 1 also reported that phonological confusions were weighted by the distance between the two confusable items. This was confirmed by comparing weighted proportions of two-apart and four-apart confusions (Footnote 2), with all 10 conditions showing a greater proportion of the former, $N=10$, $p<.005$, $CI=(.01,.08)$.

Summary of Empirical Constraints

The three meta-analyses revealed a rich set of empirical constraints on serial recall from short-term memory. In summary, the nine constraints were:

1. The primacy constraint: Recall of the first item is better than the second.

2. The recency constraint: Recall of the last item is better than the penultimate item, providing there are not too many omissions towards the end of recall.

3. The locality constraint: Items transpose small distances about their correct position.

4. The (weak) fill-in constraint: If an item is not recalled up to, or on, its correct position, it is the most likely error, other than an omission, on the following position.

5. The omission constraint: Omissions increase towards the end of recall, but not necessarily through failure to recall the last item anywhere.

6. The repetition constraint: Repetitions are literally few and far between, most often representing items recalled near the start and the end of a report.

7. The protrusion constraint: An erroneous item is more likely to occur at the same position as it appeared in the previous report than is expected by chance; intrusion of the whole report is rare.

8. The interposition constraint: Interpositions between groups are more common than expected by the locality constraint, most often between middle positions of groups, and without transposition of whole groups.

9. The confusion constraint: Phonologically similarity causes confusion in retrieval of items, but not in cuing of subsequent items (though the additional errors caused by confusions may have a small effect on retrieval of subsequent items).

Comparison of Models

Without going into the full details of the three models considered above, it is worth noting how many of the empirical constraints are met by each model.

The Power Set Model meets the primacy and locality constraints. However, it has no specified mechanism to produce omissions or repetitions. It also fails to produce sufficient recency (Murdock, 1995), probably because it does not have enough fill-in (above), and it cannot meet the confusion constraint (Henson et al., 1996). Most importantly, being a chaining model, it offers no account of the positional errors required by the protrusion and interposition constraints. These failures remain true of other variations of serial order in TODAM, such as the nesting or chunking model (Murdock, 1983, 1993, 1995).

The Articulatory Loop Model meets the primacy, locality, omission and recency constraints, though its recency is often insufficient (Burgess & Hitch, 1992). However, it does not meet the fill-in, repetition or confusion constraints (Henson et al., 1996). Being a positional model, it has the potential to meet the protrusion and interposition constraints, as demonstrated by more recent developments of the model (Burgess & Hitch, 1996a). Further revisions of the model also address the fill-in and confusion constraints (Burgess & Hitch, 1996b), though not necessarily at a quantitative level (Chapter 5).

The Primacy Model meets the primacy, recency and locality constraints (Page & Norris, 1996b), though its fill-in property is too strong (above). It also meets the omission, repetition and confusion constraints (Henson et al., 1996), though not completely satisfactorily in the case of the omission constraint (Chapter 5). Being an ordinal model however, it cannot meet the protrusion and interposition constraints.

Chapter Summary

This chapter described three meta-analyses of a number of experiments on serial recall from short-term memory. These analyses were driven by consideration of three specific models of serial recall, which make different predictions about the exact distribution of errors. They also served to confirm the generality of results in Experiments 1-3. The results of the meta-analyses were summarised in nine empirical constraints and none of the three models is able to meet all these constraints. In the next chapter, a new model is developed that can.