British Journal of Psychology (2000), **91**, 493-511 *Printed in Great Britain*

When further learning fails: Stability and change following repeated presentation of text

Catherine O. Fritz*
Psychology, Bolton Institute, Bolton, UK

Peter E. Morris
Psychology Department, Lancaster University, UK

Robert A. Bjork, Rochel Gelman and **Thomas D. Wickens**
Psychology Department, University of California, Los Angeles, USA

Kay (1955) presented a text passage to participants on a weekly basis and found that most errors and omissions in recall persisted despite repeated re-presentation of the text. Experiment 1 replicated and extended Kay's original research, demonstrating that after a first recall attempt there was very little evidence of further learning, whether measured in terms of further acquisition or error correction, over three more presentations of the text passages. Varying the schedule of presentations and tests had little effect, although performance was better when intermediate trials included both presentation and test than when only presentations or tests occurred. Experiment 2 explored whether this 'failure of further learning' effect could be overcome by (a) warning participants against basing their recall on their previous recall efforts and specifically directing them to base their recall upon the passages, (b) making each presentation more distinctive, or (c) drawing participants' attention to areas that would benefit from further learning by requiring them to tally their omissions and errors. The effect persisted in all cases. The findings have serious implications for the learning of text material.

Bartlett (1932), in his classic research on the recall of stories, demonstrated that errors and omissions, once made, persisted through repeated reproductions of his stories. He observed that the most general characteristic of repeated reproductions was the persistence of the 'form' of the first reproduction. Perhaps it is not surprising that participants tend to be consistent in their recall of a passage that they have encountered only once because they have no opportunity to refer to the correct version of the passage and so correct errors and expand their recall. However, far less well known than Bartlett's research is the research of Kay (1955), who found that errors and omissions in recall persist even with repeated presentations of the original prose passage.

Kay's (1955) participants initially heard a number of text passages and then

* Requests for reprints should be addressed to Dr Catherine O. Fritz, Psychology, Bolton Institute, Deane Road, Bolton BL3 5AB, UK (e-mail: cofl@bolton.ac.uk).

attempted to reproduce each of the first two passages. Additional presentations of the original texts followed each test. Participants were retested at weekly intervals for 5 weeks following the initial session and again 4 months after the sixth session, with each test being followed by another presentation of the passages.

Whether analysed in terms of the number of words accurately reproduced or of the meaningful content of the reproductions, Kay (1955) found that each subsequent reproduction was more like the preceding reproduction than it was like the original passage. Correct responses increased very slightly; errors and intrusions showed even smaller decreases. Essentially, participants 'were repeating the same correct items, making the same errors, and omitting the same items one week after another.... The initial reproduction had established the content of the passage; as it was laid down then so, with only slight modifications, was it held six weeks and ultimately six months later' (Kay, 1955, p. 89). In a similar procedure, spanning three weeks with four reproductions, Howe (1970) found the same pattern of results. The probability of correcting an error, once made, was very low. Whatever was produced on the first test was most likely to be produced on subsequent tests.

Intuitively, it seems strange that a learner will be less likely to produce a new correct response after hearing the information on four occasions than after hearing it once. And yet, that is exactly what occurred in Kay's (1955) and Howe's (1970) studies. Although some of the difference may be attributed to selection effects (i.e. the easiest information being learned first), something more seems to be involved. If an error was made initially, it tended to persist. Howe (1970) reported that after hearing the correct text four times, and making an error three times, the probability of producing the correct information was only .19, whereas the probability of producing a correct response on the initial test, after only a single presentation, was .43.

These data are reminiscent of a pattern of results often observed in science and mathematics learning. Students enter these domains with preconceived notions, 'naive theories', or misconceptions (e.g. Duit, 1991; Hartnett & Gelman, 1998; Kargbo, Hobbs, & Erickson, 1980; McCloskey, 1983). Like Bartlett's (1932) schemas, this prior knowledge plays a major role in the perceptions and responses that these students will construct. The errors engendered by these misconceptions, like the errors in Kay's (1955) and Howe's (1970) studies, are very resistant to correction.

If the results reported by Kay (1955) and Howe (1970) are generalizable beyond the particular schedules of presentation and testing used in their studies, then important questions are raised by their findings for education and learning in general. Is the learning process fighting against a strong tendency to remember some schema formed when information is first encountered, a schema that includes errors and omissions? If so, are there ways that this inertia can be reduced? For example, is the form of the recall fixed primarily at the time of encountering the passage or when recall is first attempted? Answers to this and related questions could identify better strategies for learning.

Kay's (1955) and Howe's (1970) presentation and test schedule resembles common educational procedures, where feedback is provided shortly after a test with further testing after a delay. Is there something about this particular sequence that reduces the effectiveness of the re-presentation? Given the proximity of the re-presentation

to the recall attempt, it might be expected that the participants would have a good opportunity to notice their errors and correct them. On the other hand, perhaps the motivation to attend may be low when recall has just been attempted and will not be required again for a week or more. Because Kay and Howe used the same presentation and test schedule, it is important to identify whether their results generalize to other schedules of presenting and testing meaningful prose passages. Experiment 1 addressed this question by manipulating aspects of the presentation and test schedules in order to provide insights into the effects of presentations, tests and arrangement.

Experiment 2 was designed to test three hypotheses as to why additional presentations fail to improve performance. One hypothesis was that the learners often report information recalled from the previous test(s) rather than from the previous presentation(s) either because the presentations of the material are less memorable than the test episodes, or they are equally memorable, but confusable. In either case, if the presentations were somehow made more distinctive, available and memorable, then further learning might be enhanced. The second hypothesis is that learners are quite capable of remembering the presentation(s), but when faced with the task of recalling the information they refer back to their previous production as an easier way to accomplish the goal. The third hypothesis is that learners fail to notice their errors and omissions and so do not correct them.

EXPERIMENT 1

The purposes of Expt 1 were to replicate Kay's (1955) and Howe's (1970) finding of persistence of omissions and errors and to identify the roles played by repeated testing, repeated presentations, and the relative order of tests and presentations in the recall of expository materials. In this latter effort, Kay's and Howe's procedure of weekly tests, each followed by presentations, was contrasted with conditions in which (a) weekly tests were preceded by presentations, (b) weekly tests were administered without additional presentations, and (c) weekly presentations occurred without tests. The first of the new conditions reduced the delay between presentation and testing. By so doing it reduced the possibility identified for the Kay schedule that participants would not attend to a presentation that would not be tested for a further week. It was also possible that the delay of a week between presentation and test in the Kay and Howe studies may have made difficult the discrimination of the presentation and the participant's previous recall attempt. In the new condition the latest presentation of the text rather than the participant's previous recall attempt was expected to be far more salient and easier to recall. Although better immediate recall was to be expected when the test followed the presentation in the same session rather than after a week, the criterion test which was used to compare the Kay/Howe procedure with the new schedule occurred in a week in which no presentation was made.

The presentation-only and the test-only conditions were introduced in an attempt to evaluate the separate contributions of presentations and tests. Performance among conditions was compared based on the fourth week when only tests were made. The inclusion of the presentation-only condition also made it possible to test whether the 'fixing' of the version of the passage that participants then repeatedly reproduce

takes place when the passage is first heard and comprehended or at the time of the first recall attempt. If the latter were the case, then a good strategy would be to delay recall while studying. However, such a strategy would need to be balanced against the known benefits of self-testing retrieval practice (Bjork, 1975).

The test-only condition provided a useful baseline against which to evaluate the contributions of the presentations. It could be argued that any benefits of learning from the presentations would be counterbalanced by forgetting during the period before the next test. Some forgetting does, undoubtedly, occur and the test-only condition measured the benefits of the tests for retention, independent of the contribution of the presentations. However, it is difficult to ascribe the results reported by Kay (1955) and Howe (1970) to forgetting. Rather, it is the consistency of recall—the remembering across the intervening weeks of the same information and the same errors—that is of particular interest.

Method

Participants

Participants were students enrolled in an introductory psychology course at the University of California, Los Angeles. They participated as one method of fulfilling a course requirement.

Eleven participants were assigned to each of four counterbalancing groups. Two participants failed to complete the experiment, leaving the groups with 11, 11, 10, and 10 participants each. Of these 42 participants, 11 were men and 31 were women. Their ages ranged from 17 to 20 (mean age 18.4).

Materials

The information to be learned. Expository text passages, approximately 200 words in length and rich in facts, were presented using audio cassette tape. Nine passages were used. Of these, four (on air traffic control, architectural mouldings, the patent process, and commercial uses of cellulose) were the experimental passages. The other passages (on garden soil, history of photography, producing watermarks, wind-generated energy, and fire extinguishers) were filler passages. The air traffic control passage is reproduced in the Appendix as an example. The topics were selected to contain information that the participants were unlikely to know in advance but would find of some minimal interest.

Participants were surveyed at the end of the last session on their prior familiarity with and interest in each topic. Familiarity ratings were on a 5-point scale: 1 = totally unfamiliar to 5 = expert. Interest ratings were also on a 5-point scale: 1 = no interest whatsoever to 5 = extremely interested. For the experimental topics, the mean familiarity rating was 1.6 and the mean interest rating was 2.3.

Design implementation

The presentation and testing of the passages was varied within participants. There were four study conditions: Test followed by Presentation (TP), Presentation followed by Test (PT), Presentation only (P), and Test only (T). The schedule associated with each condition is illustrated in Table 1. For each participant there were four experimental passages, one in each condition. To ensure that order effects and effects of specific passages were not confounded with those resulting from the assigned study conditions, study condition and initial order of presentation were assigned to passages across groups using a Greco-Latin square arrangement. Within each of the four participant groups, the passage and study condition remained linked for the duration of the experiment. The order of the passages (and thus, the study conditions) across time (in the second, third, and fourth weeks) was determined by a further Latin square arrangement. Finally, an additional Latin square arrangement was used to determine the presentation and test order of four filler passages. These filler passages were used to create controlled delays between initial presentation and test. Participants were asked to recall one of the filler passages at the end of each week's session.

Table 1. Procedural diagram of the four study conditions in Expt 1

	Week			
Study conditions	1	2	3	4
TP	P–t–TP	TP	TP	T
PT	P–t–PT	PT	PT	T
P	P–t–P	P	P	T
T	P–t–T	T	T	T

Note. P is an auditory presentation of the passage; –t– is a recall test of a filler passage; T is a recall test of the passage.

Procedure

Each week participants heard and/or were asked to write all that they could recall from each of the four experimental passages. Participants were given recorded instructions based closely upon Howe (1970) including the following: 'Please attempt to reproduce the passage about ——. Concentrate on reproducing the meaningful content of the text. Reproduce the form and phrasing as accurately as possible as well, but the primary goal is to recall the substantive content.'

Presentations took roughly 2 minutes per passage. In tests, participants were always allowed at least 2 minutes to write their recall of a passage with additional time provided as needed.

In the initial session, the nine passages were presented to the participants. The first two passages presented were filler passages, followed by an experimental passage. These presentations were followed by a test of the first filler passage and then by the appropriate activities for the experimental passage: a test and re-presentation (TP), a re-presentation and test (PT), a test (T), or a re-presentation (P). A pair of presentations (filler followed by experimental passage) occurred next, followed by a test of a filler passage (either the first or second filler passage presented) and the appropriate activities for the new experimental passage. This latter pattern was repeated twice more, so that all nine passages were presented; the first two filler passages were each tested twice; and the four experimental passages were tested and re-presented as dictated by their conditions. Participants were not aware of the distinction between filler and experimental passages. The order of the study conditions and of all passages except the first one varied across the four groups.

In the second and third weeks each of the experimental passages was presented and/or tested as dictated by its condition (see Table 1); the order of these passages varied across weeks according to the Latin square described previously. Filler passages were not re-presented, but each session ended with a test of one of the filler passages. The final week began with recall tests of each of the four experimental passages followed by a test of a filler passage, and a questionnaire.

Results

Scoring methods

Recall was scored in terms of items, a set of items being identified for each passage. Kay's (1955) use of words as the units was deemed not suitable because a report might contain correct words imbedded within an incorrect context or meaning. Scoring in terms of propositions was rejected because although propositions accurately capture meaning, they are often quite complex, containing too many elements for the proposition to be clearly identified as correct or incorrect. The items used were derived from the 'meaning units' developed by Henderson (1903; see also, Cofer, 1941; Woodworth, 1938). The scoring scheme for part of the air traffic control passage appears in the Appendix.

The participants' reports were coded with respect to the state of each item, where the states were defined as correct, incorrect, and missing. A correct item was one which was either verbatim or a reasonable paraphrase of the original text. When a report included information that did not stem directly from an item in the original passage, it was coded as an addition but these occurred too rarely for any meaningful analysis. Extracts from two recall efforts are provided, along with their codes, in the Appendix. Because the passages differed slightly in the total number of items, for each report the number of items in each state was converted to percentages of the totals and these percentages were used for analysis.

Recall performance

Participants recalled items from throughout the passages, as shown in Fig. 1. Most participants recalled information from some, but not all, parts of the passage. Almost every item was recalled by at least one person and no item was initially recalled by more than one-third of the participants.

Figure 2 shows correct recall performance across the four weeks of the experiment. The initial level of performance was somewhat low in all conditions but was within our expectations given the considerable demands of the task (i.e. learning nine text passages of 200 words each). In Week 1, as expected, the benefit of a second presentation immediately prior to the test produced better recall for the PT condition than for the TP condition; the result of a planned comparison between the two conditions was significant, $F(1,41) = 12.40$, $p < .01$. This difference was expected as the PT condition at the time of the first test had the benefit of two separate presentations with no delay between the most recent presentation and the test. For the TP condition, the second presentation occurred immediately following the test. In contrast, Week 1 tests for the TP and T conditions occurred under exactly the same circumstances with the test following one presentation with a filled interval. Performance would, therefore, be expected to be the same in Week 1, but the observed difference was statistically significant as tested by a planned comparison, $F(1,41) = 5.69$, $p < .05$. This initial, random difference needs to be taken into account because the Week 1 performance provides a baseline for the comparison of later weeks. Therefore, to simplify the visual comparison among the conditions an additional line (T adj) has been included in Fig. 2. The data points for T adj were calculated by multiplying the T values for weeks 1–4 by the TP/T ratio from Week 1.

An analysis of variance (ANOVA) was run to examine the effects of three conditions (TP, PT, and T) across the four weeks of the experiment. Because the study conditions involved different schedules for study and test it was not surprising that the condition-by-week interaction was statistically significant, $F(6,246) = 21.23$, $p < .01$. The two conditions that combined testing and re-presentation (TP and PT) showed a slight improvement from week to week, with each showing reduced performance in the first week when the test was not preceded by a presentation in that session (i.e. Week 2 for TP and Week 4 for PT). In the T condition performance showed a similar drop in Week 2 followed by only small losses in subsequent weeks.

On the criterion test in Week 4, when learning was assessed by testing all conditions without the benefit of further presentations, a main effect of study

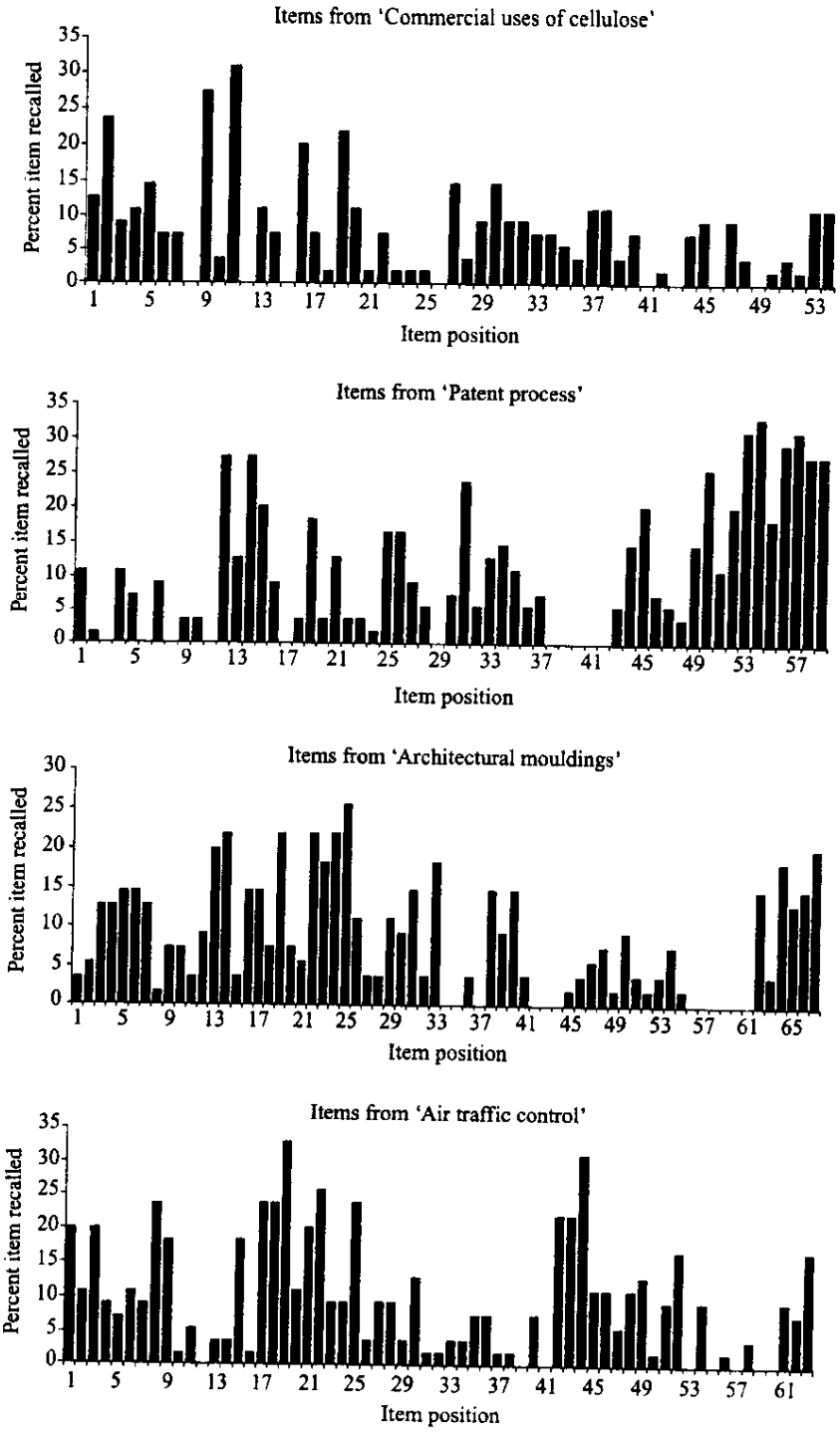

Figure 1. Percentage of participants recalling items in Week 1 in Expt 1, plotted by serial position of each item in each of the four texts.

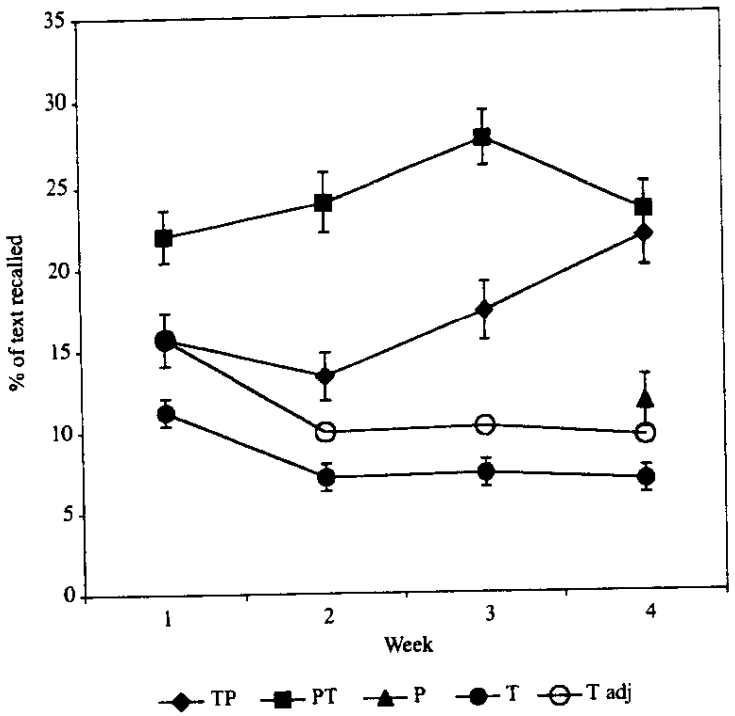

Figure 2. Percentage correct recall in Expt 1 across weeks for the study conditions (TP = Test–Present, PT = Present–Test, P = Present only, T = Test only, T adj = Test only adjusted for Week 1 performance). Error bars indicate standard errors.

condition, $F(3,123) = 49.38$, $p < .01$, was observed and tested in a one-way ANOVA. Tukey tests identify that the PT and TP conditions were not reliably different but the P condition was reliably poorer than both the PT and TP conditions ($p < .05$) and the unadjusted T condition was reliably poorer than the other three conditions ($p < .05$).

No significant effects or interactions were found for production of incorrect items analysed across weeks and conditions.

Stability and change

The design made possible the comparison of the fate of items in the PT, TP and T conditions between the first test and the final, criterion test. Changes between the first and final recall attempts were classified for all possible combinations: correct to correct, correct to incorrect, and so forth. The percentage of items in each situation, combining the PT and TP conditions, are given in Table 2.[1]

[1] The TP and PT conditions were combined for clarity. Tukey tests following a MANOVA analysis for differences between the conditions found a significant difference between PT and TP for only one measure: there were more correct items that remained correct in the PT condition ($p < .05$). This difference is an artifact of the different performance levels in Week 1. Analysis of the probability for correct items to remain correct, when conditionalized upon initial correct performance, found no significant difference.

Table 2. Experiment 1: State of items in Week 1 and Week 4 as a percentage of total items

State of item in Week 1	State of item in Week 4			Week 1 Total
	Correct	Incorrect	Missing	
PT and TP conditions combined				
Correct	12.6	0.8	5.4	18.8
Incorrect	1.3	2.2	1.7	5.2
Missing	8.7	2.0	65.3	76.0
Week 4 Total	22.6	5.0	72.4	
T condition				
Correct	4.8	0.9	5.4	11.1
Incorrect	0.3	2.0	1.7	4.0
Missing	1.7	0.5	82.7	84.9
Week 4 Total	6.8	3.4	89.8	

Despite the repeated presentations of the original text in each week between the first and final tests, in the TP and PT conditions 80% of the items were unaffected across trials. The majority of these were items that were missed on the first test and were still missing on the final test. However, although almost 19% of the items were correct on the first test; only a further 4% were correct on the final trial.

The following results refer to the patterns of data in Table 2 and report additional probabilities that are conditional upon the initial state of the items. For the PT and TP conditions, 18.8% of the total items were reported as correct in Week 1. Of those items, 67% (or 12.6% of the total) were also correct in Week 4. Most of the remaining items were lost (28%, or about 3.5 items on average). Very few items were altered to become incorrect (only about 0.5 of an item on average). Initial performance was poorer for the T condition: only 11.1% of the items were reported correctly in Week 1. Fewer than half of these were also reported correctly in Week 4 whereas slightly more than half of these items were missing from the final report.

Items incorrect in Week 1 were scarcer, with only about 3 items (4–5% of the total) reported incorrectly. Of these it appears that slightly more than one item remained incorrect, one erroneous item was dropped, and one was corrected for the PT and TP conditions. In the T condition the initial errors were not corrected, but otherwise the pattern was similar to the other conditions.

The items that were initially missing were most likely to remain missing. For the PT and TP conditions 76% of the total items (roughly 46 items) were missing from the Week 1 reports. Of these items, 87% ($\approx$ 40 items) were also missing from the Week 4 report; only 11% ($\approx$ 5) of these items appeared correctly in the final report following two or three additional presentations and two intervening tests. For the T condition 85% of the total items ($\approx$ 53 items) were missing from the initial reports; on average just one of those items was present and correct in the Week 4 report.

Considering the data from the opposite direction, correct performance in the Week 4 criterion test derived from initial correct performance, with some losses, and

from acquisition of a few items that had been missing in Week 1. With repeated presentations roughly one correct item in Week 4 might have initially been reported incorrectly.

Discussion

Our results replicate and extend those of Kay (1955) and Howe (1970). Like Kay and Howe we found that performance changed little after the first week. Subsequent presentations of the passages had comparatively little effect in adding new information to future recall or in correcting errors. For the conditions that involved both repeated presentations and tests there was a net gain of less than 4% of the total items correct at the final test as compared to the first test. Even after several additional presentations and tests, an error was almost twice as likely to be repeated as to be corrected.

A possible explanation for the phenomenon might be that the items making up the passages differ greatly in the ease with which they can be learned. Participants might be acquiring most of the easy items on the first trial, leaving only difficult items for subsequent trials. Reliance upon easy items, and avoidance of difficult items, would produce the observed consistency across trials. This explanation is unconvincing given the pattern of recall on the first trial. This pattern is illustrated by the frequency with which items were remembered for every item in each of the four passages (see Fig. 1). The pattern is inconsistent with an 'easy item' explanation. Most items were recalled by at least one participant, and no items were recalled by more than a third of the participants. To the degree that some items are generally easier or more difficult than others, it is clear that easy items remained available for all participants after the first trial. Furthermore, we observed that most participants' recall was limited to parts of the passage. Because easy items are expected to appear throughout the passage, most participants had a ready pool of easy items available. In any case, 81% of items remained to be acquired correctly after the first test suggesting that there were ample opportunities for further items to have been learned.

It is clear that both tests and additional presentations made a contribution to performance: the PT and TP conditions produced better performance in Week 4 than either the P or T condition. Analysis of the fate of individual items in the PT and TP conditions as compared with the T condition discloses that additional presentations both helped to maintain initially correct items and to acquire a few additional items. Although the combination of presentation and testing was a more effective arrangement than tests or presentations alone, the benefits observed were relatively small.

Why do participants learn so little further information after their first recall attempt? We address several possible hypotheses in Expt 2, but some alternatives can be dismissed on the basis of the present results. Because the TP condition did not differ on the final trial from the PT condition we conclude that the results reported by Kay (1955) and Howe (1970) are not specific to one pattern of presentation and testing of the correct passages. Despite being tested immediately following a fresh, complete representation of the material during learning trials, the PT condition failed to yield better performance than the TP condition in the criterion test. The T condition provides an answer to the concern that forgetting over the weeks might severely influence performance. Although there is a decline in the T group recall

between the first and second tests, there is no noticeable decline thereafter. The implication is that some information is lost between the first and second tests and this loss needs to be kept in mind when interpreting the stability and change data. However, the size of this decline is small in comparison with the amount that is acquired on the first trial. Forgetting between trials does not seem to be the explanation of the failure of further leaning. In any case, as noted earlier, the most interesting aspect of the data is their stability or, in other words, the amount that is remembered but not expanded or corrected.

This experiment demonstrated that the phenomenon reported by Kay (1955) and Howe (1970), which we call the failure-of-further-learning effect, is robust across testing conditions. The very slow rate of acquisition of new information after the first encounter with the passage is in marked contrast to traditional list-learning experiments where steady acquisition across trials is the norm (e.g. Taylor & Irion, 1964; Underwood, Runquist, & Schulz, 1959). In Expt 2 we explored three hypotheses that might account for the failure-of-further-learning effect.

EXPERIMENT 2

Experiment 1 demonstrated how repeated presentations of meaningful text led to very little acquisition after the first trial, with most errors remaining uncorrected. Why should so little be acquired from these subsequent presentations and why are errors resistant to correction? In our second experiment we investigated three hypotheses about learners' behaviour that might contribute to this failure of further learning.

The first hypothesis suggests that, in tests after the first one, learners base their recall on their memory of previous tests, rather than previous presentations, either because they cannot remember the presentations or because they fail to attribute correctly the source of conflicting information that comes to mind. Given that generated information tends to be better remembered than presented information (Jacoby, 1978; Slamecka & Graf, 1978; Wittrock, 1974) and that testing has a powerful effect upon subsequent recall (Bjork, 1975) this hypothesis seemed worth exploring. We tested this hypothesis by seeking, in one condition, to make the original presentations more distinctive and eventful to increase their memorability (Hunt & McDaniel, 1993).

The second hypothesis is that memories of the presentations and the test may both be accessible and distinguishable, but that learners' efforts to recall refer back to their previous report(s), rather than the prior presentation(s), because doing so is easier—perhaps because the new recall task resembles the earlier recall tasks more than it does the presentations. According to this hypothesis, recall omissions and errors result from a strategic decision by the participants rather than a limitation on what can be recalled. Lansdale and Laming (1995) reported that their participants repeated errors in their attempts to recall the details of objects presented on a billiard table. They attributed the errors to the participants finding it easier to recall a previous report than to attempt a new one. Although the recall of text may differ from that of objects, a similar approach by participants may be operating. We tested this possibility with a condition that included instructions that emphasized recalling from the presentations and not from previous recall attempts.

The third hypothesis is that, during the presentations that follow a test, learners fail to notice information that was omitted from or is inconsistent with their previous reports. If the incorporation of new information or error correction require an active changing of an established schema, then to identify where changes are required participants would have to remember their own version while listening to the re-presentation of the passage so that they could identify inconsistencies. Because this triple task (remembering, listening, and modifying) may not be easy it is possible that participants do not, under normal conditions, engage in such a process. As a test of this hypothesis, we included a condition in which participants were explicitly asked, during re-presentations, to tally omissions and errors from their previous recall(s).

As in Expt 1, participants were asked to learn expository text passages over a 4-week period. All tests and presentations were scheduled in the Test-Present sequence, matching Kay's (1955) and Howe's (1970) procedures, but instructions and some details of presentation were manipulated between participants.

Method

Design

There were four conditions of study, with one group of participants assigned to each condition: a Control group without special instructions or presentations; a Directed group with special instructions for the tests; an Eventful group with special presentations; and a Notice group with special instructions for the presentations. Each participant attempted to learn three passages. The order of the passages was varied across Weeks 1–3 by a Latin square; in Week 4 the passages were tested in the order used in Week 1. Within each week the passages were presented and tested in the same order for all participants.

Participants

Participants were recruited from the same pool as for Expt 1.

Experiment 2 began with 62 participants, divided among the four experimental groups. Five participants either failed to reappear or failed to follow instructions correctly during the sessions and were dropped from the experiment. The final group sizes were Control 14, Directed 12, Eventful 16, and Notice 15. Of those completing the experiment, 18 were men and 39 were women. Their ages ranged from 17 to 21 (mean age 18.26).

Materials

Three of the passages from Expt 1 (architectural mouldings, air traffic control, and the patent process) were used as the experimental passages and a fourth passage (history of photography) was presented as an orienting task but not tested.

Videotaped instructions and presentations. In the first three sessions, instructions and expository passages were read to the participants by a person on videotape. In an effort to create more distinctive, memorable presentations for the Eventful condition, four different readers were videotaped: a young man, a young woman, a more mature (50s) man and a more mature (50s) woman. The settings for the four readers varied. The young man's head and upper body was shown in an office; the young woman's head before a pale blank wall; the middle-aged man was shown from the chest up, outdoors, in front of a fence; and the middle-aged woman's head and shoulders were shown in front of a living room wall bearing a wire sculpture. For the Control, Directed and Notice conditions, each presentation of every passage was read by the same person: the young woman. For the Eventful condition the young woman read the passage for the first two (pre-test) presentations but the third presentation introduced another reader, as did the fourth and the fifth.

Error tally pages. For the Notice group, small (roughly 13.5 × 10 cm) pages were prepared on which the participants tallied in separate columns the errors and omissions that they observed while listening to the passage being re-presented.

Mathematics booklets. A booklet of basic mathematics problems involving arithmetic with fractions was used as an activity to fill scheduled delays.

Procedure

Once the participants were assembled, a pre-recorded introduction and initial instructions were presented via the videotape. The basic sequence of events was the same for all participants, although specific instructions and concurrent activities differed between groups. An initial passage about the history of photography was presented as an orientation activity but was never tested. Then, one of the experimental passages was presented twice, followed by a 3-minute delay during which the participants completed mathematical problems. Their recall of the passage was then tested and the passage was presented a further time. The next experimental passage was then presented twice, followed by a 3-minute interval of mathematical problems, a test of that passage and its re-presentation. The same procedure was then followed for the remaining experimental passage.

The instructions for the initial presentation, including the photography passage, were: 'Please listen carefully to the following passage about ——. Try to understand and remember the passage, as you may be tested on it at some later time.' The standard instructions for all subsequent presentations in all sessions were: 'Please listen once again to the passage about ——. Try to understand and remember the passage, as you may be tested on it at some later time.'

Following the two initial presentations and the filled interval, participants were instructed to recall the passage. The instructions included: 'Concentrate on reproducing the meaningful content of the passage. Reproduce the form and phrasing of the original passage as accurately as possible as well, but the primary goal is to recall the meaningful content.'

In the second and third weeks, the passages were presented in different orders. The procedure included the interval filled with mathematical problems, the test and the re-presentation of each passage. The instructions were identical to those in Week 1.

The recall tests in the fourth week constituted a criterion test. These tests were conducted without intervening delays or presentations.

The experimental manipulations. For the Control group, the instructions were always the standard instructions, described above. The other conditions were variations on the Control condition. The Directed group was like the Control group except that in Weeks 2 and 3 they were explicitly advised to recall the previous *presentations* and not their previous reports. The standard test instructions were given to the Directed group with the following addition.

> As you try to recall the passage, be sure to think back to when you *heard* the passage, not to when you wrote it! People who think back to what they have written before make the same errors repeatedly. *You* can avoid this problem by being careful to remember what you *heard* rather than what you wrote. (Verbal emphasis was obvious.)

The instructions for the Eventful group did not differ from the Control group, but the actual presentations differed, not in content, but in reader. This group saw four different readers during the experiment. In the first session one reader was used for the initial two presentations (the young woman) and another reader was used for the presentation following the test (the mature man). The reader for the second week was the young man and the reader for the third week was the mature woman.

The Notice group was similar to the Control group except that for each presentation following a test the participants were advised to notice any omissions or errors that they had made in that test. Participants were given the omissions and errors page and were asked to make a tally mark in the appropriate column whenever they noticed, while listening to the passage, that they had omitted some information from the previous test or made an error. New pages were provided for each presentation following each test. The instructions for this task, presented just before the standard instructions for a subsequent presentation, were:

> You are about to hear the passage about —— again. As the videotape presents the passage again, please pay careful attention to any errors you might have made and to any parts that you may have

left out. Whenever you notice that you made an error, something that you wrote that was wrong, please make a tally mark under the 'errors' heading. Whenever you notice something that you left out of your reproduction, make a tally mark under the 'omissions' heading.

While listening to these instructions, participants saw an example of an omissions and errors page. The first time that these instructions were presented, they included specific references to the example on the videotape.

Results

The scoring methods and measures developed for Expt 1 were also used for Expt 2. Performance across the 4 weeks is illustrated in Fig. 3.

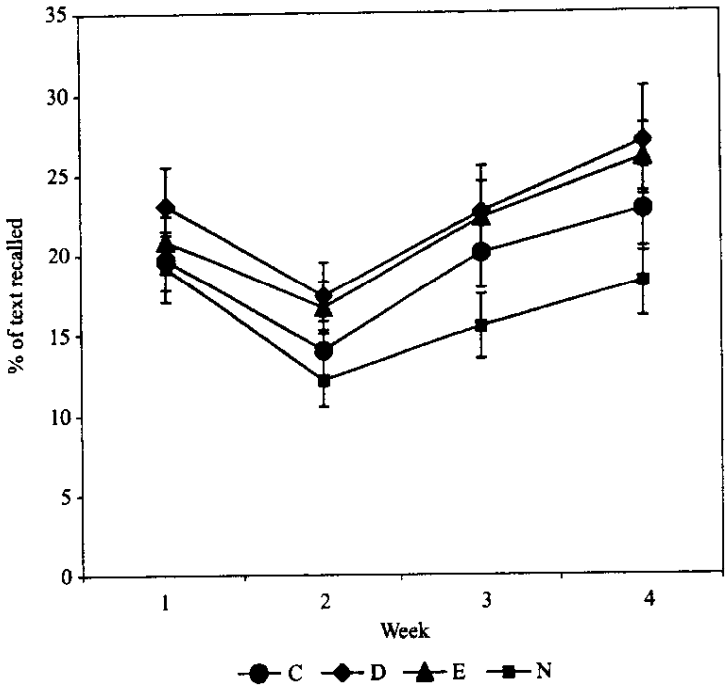

Figure 3. Percentage correct recall in Expt 2 across weeks for the study conditions (C = Control, D = Directed, E = Eventful, N = Notice). Error bars indicate standard errors.

Recall performance

Prior to the first test in Week 1, all participants had the same experience with each tested passage, that is, two presentations, and the same instructions with respect to that passage. Therefore, no differences in performance were expected in Week 1 and none were observed, $F(3,53) < 1$.

The level of performance was, as intended, slightly higher than in Expt 1. The pattern across the weeks replicated that of the TP condition in Expt 1, including a drop for Week 2 followed by slight improvement in subsequent weeks. A condition by week interaction was observed and tested to be significant in a two-way (condition × week) analysis of variance (ANOVA), $F(9,159) = 2.21$, $p < .05$. The locus of the interaction was the different slope for the Notice condition compared to the other conditions. Dropping the Notice data from the analysis removed the

interaction, $F(6,118) < 1$. By Week 4, the Notice group were performing significantly more poorly than the other groups, $F(1,53) = 4.21$, $p < .05$.

Table 3. Experiment 2: State of items in Week 1 and Week 4 as a percentage of total items

State of item in Week 1	State of item in Week 4			Week 1 Total
	Correct	Incorrect	Missing	
Correct	14.4	1.5	5.0	20.9
Incorrect	2.0	2.8	2.0	6.8
Missing	7.5	2.5	62.4	72.4
Week 4 Total	23.9	6.8	69.4	

Stability and change

As for Expt 1, the fate of items that had been correct, incorrect or missing in Week 1 was examined and is reported in Table 3. Because a one-way MANOVA analysis of the four conditions yielded no significant differences on any of the change measures, data in the table are collapsed across all conditions. The patterns of stability and change that occurred were very similar to those occurring in Expt 1, as can be seen by comparing Table 2 and Table 3. The vast majority (80%) of items did not change their state over the four trials despite the repeated presentations of the text. In Week 1, 21% of the items were reported correctly; for Week 4 the net gain was only 3%.

Following the pattern in Expt 1, 69% of the items reported correctly in Week 1 (or 14% of the total) were also correct in Week 4, with most of the remaining items dropping out of the reports (24%). Few items were initially reported incorrectly (only 7% of total items, roughly four items per passage); as in Expt 1, roughly one-third of these were corrected, one-third persisted and one-third did not appear in the final report. In Week 1 72% (roughly 45 items) were missing; roughly one-third of those were correctly reported in Week 4 with most of the rest still missing. The items added to the correct report (9.5% of the total) represent a gain of less than half of the items initially reported correctly (20.9%).

Discussion

It appears that none of the three hypotheses explored in Expt 2 accounts for the failure of further learning observed by Kay (1955) and Howe (1970) and replicated here. Directing participants' attention to recalling the passages as presented rather than as previously recalled did not help, nor did increasing the distinctiveness of each presentation in the Eventful condition. The failure-of-further-learning effect appears to be robust and not easily overcome either by instructions to the participants or by increasing the distinctiveness of the presentations. We also observed that drawing the attention of the participants to their omissions and errors, in the Notice condition, led to *poorer* recall, not better recall. This may have been because the task of

evaluating memories of the passage and noting omissions and errors made attending to acquiring new information from the passage more difficult, or perhaps because an emphasis on error avoidance shifted some response criterion or increased processing of prior errors.

The stability of the recall after the first presentation was once again evident, despite our efforts to improve performance. On the first trial 21% of items were recalled correctly and only 24% were recalled correctly after three further presentations. After the additional presentations, the number of errors was unchanged.

GENERAL DISCUSSION

In our first experiment we replicated and extended earlier work by Kay (1955) and Howe (1970), who found that initial recall performance for text materials persisted, even in the face of additional learning trials. We found that after the first recall attempt participants' recall was surprisingly resistant to change. Most of the information that was missing from the first recall attempt remained missing after several more presentations and tests. Errors made in the first recall attempt were more likely to be retained rather than be corrected.

One potential explanation of the failure of further learning—that the failure was the result of the scheduling of the presentations—was ruled out. Experiment 2 tested three possible explanations for the effect, but found no support for any of them. There are other possible explanations.

It may be that the familiarity of a passage when it is encountered for a second or subsequent time in some way inhibits further learning. Perhaps with connected, meaningful text a schema of the text is formed when it is first encountered and this schema is resistant to change. These experiments intentionally avoided narrative texts to avoid the schematizing effects of story grammars (e.g. Mandler & Johnson, 1977). It is possible, though, that people quickly develop and apply schemas to expository text as well. Indeed, this general sort of explanation is consistent with the persistence of naïve theories in the face of science instruction (e.g. Duit, 1991; Hartnett & Gelman, 1998; Kargbo *et al.*, 1980; McCloskey, 1983).

It may also be that text that has been encountered previously is processed in a different way when its familiarity is recognized. Clark and Haviland (1977) have suggested that the information that we encounter is analysed for what is 'given' from previous experience and what is 'new'. The latter receives processing attention and the former is relatively neglected. When a previously heard text passage is recognized as having been heard before it could be classified as 'given' and therefore receive relatively little attention. Because what we remember is very much a by-product of our perceptual and comprehension processes (Bransford & Johnson, 1973 ; Craik & Tulving, 1975), what is already familiar receives much less processing than what we consider new. The consequence is that less is acquired from information that is perceived as already known, even though the knowledge acquired may be sufficient only for general recognition and not for recall.

It is worth noting again that the failure of further learning that we have observed differs from the steady increments in learning that are normally reported when lists of items are repeatedly presented. The explanation for this difference may lie in the

coherent nature of the textual material that we have been investigating but may also be influenced by the spacing of the re-presentations in our study. Traditional list-learning studies normally repeat lists after minutes rather than the weekly intervals used in the present studies. However, the spacing of the presentations and tests should have helped rather than hindered learning insofar as spaced practice is normally found to be more beneficial than massed practice (e.g. Dempster, 1996).

So far as we can see, existing accounts of human memory do not predict the failure of further learning. Certainly, it must be possible to tune models to produce a reasonable level of initial learning followed by a drastically reduced level of further learning, but fitting a model to these data does not explain the data. Is the failure of further learning limited to some particular levels of initial performance? Does it apply to learning of different sorts of materials? (Note that Kay (1955) first observed it in a sequence learning task—rather different from the text materials that have been used since.) Is the failure of further learning to some degree a function of learners' expectations about their performance? Is there a more general principle involved—one that people might apply to performance in non-learning situations as well, such as the 80/20 rule (80% of the results come from 20% of the effort)? Greater understanding of the failure of further learning may lead to a greater understanding of some fundamental aspects of human memory and possibly of human cognition in more general terms.

From an applied perspective, this effect may be a major impediment to learning. A glance at the science and mathematics learning literature assures us that the perseverance of earlier ideas over new learning is not unique to the present experiments. Finding a way of overcoming the failure of further learning is crucial for effective education and training.

Acknowledgements

The data were collected while the first author was a graduate student at UCLA supported by a National Science Foundation Graduate Research Fellowship. The authors thank Rachel Henley and two anonymous reviewers for their helpful comments on an earlier draft.

References

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bransford, J. D., & Johnson, M. K. (1973). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior, 11*, 717–726.

Clark, H. H., & Haviland, S. E. (1977). Comprehension and the given-new contract. In R. O. Feedle (Ed.), *Discourse production and comprehension* (pp. 1–40). Norwood, NJ: Ablex.

Cofer, C. N. (1941). A comparison of logical and verbatim learning of prose passages of different lengths. *American Journal of Psychology, 54*, 1–20.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*, 268–294.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Memory* (pp. 317–344). San Diego: Academic Press.

Duit, R. (1991). Students' conceptual frameworks: Consequences for learning science. In S. M. Glynn, R. H. Yearny, & B. K. Britton (Eds.), *The psychology of science learning* (pp. 65–85). Hillsdale, NJ: Erlbaum.

Hartnett, P., & Gelman, R. (1998). Early understanding of number: Paths or barriers to the construction of new understandings? *Learning and Instruction, 8*, 341–374.

Henderson, E. N. (1903). A study of memory for connected trains of thought. *Psychological Monographs, 5*(23).

Howe, M. J. A. (1970). Repeated presentation and recall of meaningful prose. *Journal of Educational Psychology, 61*, 214–219.

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language, 32*, 421–445.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17*, 649–667.

Kargbo, D. B., Hobbs, E. D., & Erickson, G. L. (1980). Children's beliefs about inherited characteristics. *Journal of Biological Education, 14*, 137–146.

Kay, H. (1955). Learning and retaining verbal material. *British Journal of Psychology, 46*, 81–100.

Lansdale, M., & Laming, D. (1995). Evaluating the fragmentation hypothesis: The analysis of errors in cued recall. *Acta Psychologica, 88*, 33–77.

Mandler, J. M., & Johnson, N. S. (1977). Remembrance of things parsed: Story structure and recall. *Cognitive Psychology, 9*, 111–151.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models* (pp. 299–324). Hillsdale, NJ: Erlbaum.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*, 597–604.

Taylor, A. B., & Irion, A. L. (1964). Continuity hypothesis and transfer of training in paired-associated learning. *Journal of Experimental Psychology, 68*, 573–577.

Underwood, B. J., Runquist, W. N., & Schulz, R. W. (1959) Response learning in paired-associated lists as a function of intralist similarity. *Journal of Experimental Psychology, 58*, 70–78.

Wittrock, M. C. (1974). Learning as a generative process. *Educational Psychologist, 11*, 87–95.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Holt.

Received 18 October 1999; revised version received 13 March 2000

Appendix

One of the text passages:
Air traffic control

Long distance travel today is almost always by air—a speedy and safe way to cross continents and oceans. The safety of air travel is due in great part to sophisticated air traffic control systems which direct planes through the airspace. The airspace is composed of airways, control zones, and control areas. Airways are precisely defined corridors that link airports; control zones cover the space within 5 miles of an airport; and control areas are broader controlled spaces near airports.

In the United States, Air Traffic Control Towers are responsible for directing planes located in control zones and control areas. Controllers keep track of the identification, intentions, direction, timing, height, and position of aircraft using a flight progress board and a plan position radar.

In poor weather and times of heavy traffic, it is often necessary to 'stack' planes prior to landing. Stacking is accomplished using a vertically directed stacking beacon, around which planes circle to starboard, that is, turning right. Levels in the stack are separated vertically by 1000 feet. An Instrument Landing System, with localizer beacons for direction, glidepath beacons for altitude, and marker beacons for distance, can be used to guide planes to a safe landing when visibility is poor.

Items defined for the first parts of 'Air traffic control'

The opening paragraph of the 'Air traffic control' passage is shown in the table in terms of items. Examples from two representative productions (by participants 'AT' and 'CMF') are included with corresponding scores.

Original text	Indicative 'item' content	Example 1, 'AT', Week 1		Example 2, 'CMF', Week 1	
		Text	Code	Text	Code
Long distance travel today is almost always by air–	Air travel			Travel through air is very common these days	c
	Long distances				
	Almost always				c
a speedy and safe way to cross continents and oceans.	crossing				
	continents				
	oceans				
	speedy				
	safe	Airplanes safety is helped by the use of air traffic control	c		
The safety of air travel is due in great part to sophisticated air traffic control systems	Safety due to ATC		c	Air traffic control helps regulate it by controlling the air in the airways, control zones, and control areas. The airways are the spaces between airports, the control zones are the five mile zones around the airports, and the control areas are the areas near the runways.	
	ATC systems				
	sophisticated				
which direct planes through the airspace.	Direct planes				
	Through airspace				
The airspace is composed of airways, control zones, and control areas. Airways are precisely defined corridors that link airports; control zones cover the space within 5 miles of an airport; and control areas are broader controlled spaces near airports.	airways	This includes air lanes, safety areas, and safety zones . . . Safety areas are within 5 miles and safety zones are a bit wider than that. But air lanes are tight.	x		c
	Precisely defined		x		
	Corridors				x
	Linking airports				c
	Control zones		x		c
	Within 5 miles		c		c
	Near airports				c
	Control areas		x		c
	Broader areas		c		
	Near airports				x