Deakin University
Disposition of Thesis

I am the author of the thesis entitled *A Behavioural Analysis of Enforced Delays in Computerised Programmed Instruction* submitted for the degree of Doctor of Philosophy and agree to the thesis being made available for such consultation, loan or photocopying as may be approved by the University Librarian, provided that no part of the thesis shall be reproduced without the prior approval of the University Librarian and with the appropriate acknowledgment of the source.

Signature Redacted by Library

Glenn Kelly
December 1995
DEAKIN UNIVERSITY LIBRARY

TO:  ALL USERS OF THIS THESIS

Please sign this form to indicate that you have used this thesis in accordance with the disposition signed by the author of this thesis.

Thank you.

SUE McKNIGHT
University Librarian

Name  Signature  Date

[Redacted Signature]  1/6/96

[Redacted Signature]  26/6/03

DEAKIN UNIVERSITY LIBRARY
C011 0797597 2
92.7.029.047.270
A Behavioural Analysis of Enforced Delays in Computerised Programmed Instruction

Glenn Kelly B.Sc. (Hons.)

This thesis is submitted in fulfilment of the requirements for the degree of Doctor of Philosophy.

Faculty of Health and Behavioural Sciences
Deakin University
December 1995
Deakin University
Candidate's Certificate

I certify that the thesis entitled *A Behavioural Analysis of Enforced Delays in Computerised Programmed Instruction* submitted for the degree of Doctor of Philosophy is the result of my own research, except where otherwise acknowledged, and that this thesis in whole or in part has not been submitted for an award including a higher degree to any other university or institution.

Signature Redacted by Library

Glenn Kelly
December 1995
# Table of contents

Declaration ........................................................................................................... ii
Table of Contents ................................................................................................... iii
Tables and Figures ................................................................................................ vi
Acknowledgments ................................................................................................. xii
Abstract ................................................................................................................ xiii

CHAPTER 1: Literature Review............................................................................. 1
  Overview.............................................................................................................. 1
  Automated Instruction....................................................................................... 1
    Pressey............................................................................................................. 2
    Skinner............................................................................................................ 4
  Influence of Skinner's Contribution................................................................. 9
    Instructional effectiveness.............................................................................. 10
    The decline of programmed instruction....................................................... 17
    Summary of Skinner's contribution............................................................... 21

Computer-Based Instruction (CBI)................................................................. 22
  Typology.......................................................................................................... 22
  Origins of CBI.................................................................................................. 24
  Effectiveness.................................................................................................... 27
  Do computers contribute anything to learning?........................................... 31
  Quality problems in CBI................................................................................. 34
  Summary and future research........................................................................ 39

Learner Control ................................................................................................. 40
  Learner control of pacing.............................................................................. 42

Experimentally-Imposed Delays...................................................................... 44
  Prerresponse delays...................................................................................... 44
  Prefeedback delays........................................................................................ 53
  Postfeedback delays...................................................................................... 61
  Delays in system response time and user response time............................ 69
  Experimentally-imposed delays: summary.................................................... 75

Rationale and the Present Studies................................................................. 76

CHAPTER 2: General Method........................................................................ 80
  Experimental Design....................................................................................... 80
    Single-subject designs.................................................................................. 80
    Interactions................................................................................................... 85
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Design Considerations for the Present Experiments</td>
<td>87</td>
</tr>
<tr>
<td>Computerised Programmed Instruction</td>
<td>88</td>
</tr>
<tr>
<td>Subjects</td>
<td>94</td>
</tr>
<tr>
<td>Apparatus</td>
<td>94</td>
</tr>
<tr>
<td>Procedure</td>
<td>95</td>
</tr>
<tr>
<td>Analyses</td>
<td>101</td>
</tr>
<tr>
<td>CHAPTER 3: Effects of Enforced Delays</td>
<td>105</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>106</td>
</tr>
<tr>
<td>Method</td>
<td>106</td>
</tr>
<tr>
<td>Results</td>
<td>107</td>
</tr>
<tr>
<td>Discussion</td>
<td>119</td>
</tr>
<tr>
<td>Experiment 2</td>
<td>122</td>
</tr>
<tr>
<td>Method</td>
<td>123</td>
</tr>
<tr>
<td>Results</td>
<td>124</td>
</tr>
<tr>
<td>Discussion</td>
<td>134</td>
</tr>
<tr>
<td>CHAPTER 4: Effect of Delays on Workrate and Retention</td>
<td>138</td>
</tr>
<tr>
<td>New Outcome Measures</td>
<td>138</td>
</tr>
<tr>
<td>Time components</td>
<td>138</td>
</tr>
<tr>
<td>Retention</td>
<td>138</td>
</tr>
<tr>
<td>Procedural Changes</td>
<td>139</td>
</tr>
<tr>
<td>Design</td>
<td>139</td>
</tr>
<tr>
<td>Review sets</td>
<td>140</td>
</tr>
<tr>
<td>Reviews</td>
<td>140</td>
</tr>
<tr>
<td>Experiment 3</td>
<td>140</td>
</tr>
<tr>
<td>Method</td>
<td>141</td>
</tr>
<tr>
<td>Results</td>
<td>142</td>
</tr>
<tr>
<td>Discussion</td>
<td>156</td>
</tr>
<tr>
<td>Experiment 4</td>
<td>162</td>
</tr>
<tr>
<td>Method</td>
<td>164</td>
</tr>
<tr>
<td>Results</td>
<td>165</td>
</tr>
<tr>
<td>Discussion</td>
<td>180</td>
</tr>
<tr>
<td>CHAPTER 5: Effects of Reviews</td>
<td>182</td>
</tr>
<tr>
<td>Experiment 5</td>
<td>194</td>
</tr>
<tr>
<td>Method</td>
<td>194</td>
</tr>
<tr>
<td>Results</td>
<td>195</td>
</tr>
<tr>
<td>Discussion</td>
<td>196</td>
</tr>
<tr>
<td>Experiment 6</td>
<td>198</td>
</tr>
<tr>
<td>Section</td>
<td>Page</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>Method</td>
<td>199</td>
</tr>
<tr>
<td>Results</td>
<td>199</td>
</tr>
<tr>
<td>Discussion</td>
<td>211</td>
</tr>
<tr>
<td>CHAPTER 6: General Discussion</td>
<td>214</td>
</tr>
<tr>
<td>Effectiveness of Delays</td>
<td>214</td>
</tr>
<tr>
<td>Activity During Delays</td>
<td>215</td>
</tr>
<tr>
<td>Motivation</td>
<td>217</td>
</tr>
<tr>
<td>Compensatory Pacing</td>
<td>220</td>
</tr>
<tr>
<td>The Self-Pacing Tenet</td>
<td>221</td>
</tr>
<tr>
<td>Duration of Delay</td>
<td>224</td>
</tr>
<tr>
<td>Reviews</td>
<td>225</td>
</tr>
<tr>
<td>Time-On-Task Notion</td>
<td>227</td>
</tr>
<tr>
<td>Experimental issues</td>
<td>229</td>
</tr>
<tr>
<td>Programmed Instruction</td>
<td>234</td>
</tr>
<tr>
<td>Future directions</td>
<td>237</td>
</tr>
<tr>
<td>APPENDICES</td>
<td>241</td>
</tr>
<tr>
<td>Appendix 2-1: Table of contents from The Analysis of Behavior</td>
<td>241</td>
</tr>
<tr>
<td>Appendix 2-2: Subject details</td>
<td>244</td>
</tr>
<tr>
<td>Appendix 2-3: Experimental instructions for Experiment 2</td>
<td>245</td>
</tr>
<tr>
<td>Appendix 2-4: Experimental instructions for Experiments 3 and 4</td>
<td>246</td>
</tr>
<tr>
<td>Appendix 2-5: Experimental instructions for Experiments 5 and 6</td>
<td>247</td>
</tr>
<tr>
<td>Appendix 2-6: Consent form for Experiments 1 and 2</td>
<td>248</td>
</tr>
<tr>
<td>Appendix 2-7: Consent form for Experiments 3 to 6</td>
<td>249</td>
</tr>
<tr>
<td>Appendix 2-8 to Appendix 2-12: Example screen displays</td>
<td>250</td>
</tr>
<tr>
<td>Appendix 2-13: Percentage of incorrectly scored responses</td>
<td>255</td>
</tr>
<tr>
<td>Appendix 3-1: Sequence of conditions in Experiment 1</td>
<td>256</td>
</tr>
<tr>
<td>Appendix 4-1: Sequence of conditions in Experiment 3</td>
<td>257</td>
</tr>
<tr>
<td>Appendix 4-2: Sequence of conditions in Experiment 4</td>
<td>258</td>
</tr>
<tr>
<td>Appendix 5-1: Sequence of conditions in Experiments 5 and 6</td>
<td>259</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>260</td>
</tr>
</tbody>
</table>
Tables and figures

CHAPTER 1: Literature Review

Table 1-1 Percentage of Studies in Which the Effect of Programmed Instruction (PI) and Traditional Instruction (TI) on Achievement was Statistically Equivalent, or Statistically Significant in Favour of One Method.................................................................15

Table 1-2 One Classification of Computer-Based Instructional Applications...............23

Table 1-3 Percentage of Studies in Which the Effect of Computer-Based Instruction (CBI) and Traditional Instruction (TI) on Achievement was Statistically Equivalent, or Statistically Significant in Favour of One Method, Plus the Mean Effect Size.................................................................29

Table 1-4 Sequence of Events in Brackbill (1964) .........................................................62

Figure 1-1 Pressey's automated teaching device..........................................................2

Figure 1-2 One version of a Skinnerian teaching machine .............................................8

Figure 1-3 Formula for calculating effect size ..............................................................14

Figure 1-4 Teaching machine and programmed instruction cartoon.............................19

Figure 1-5 Major elements in a typical learning trial and the time intervals that separate them ..............................................................................................................44

Figure 1-6 A typical sequence of events and location of delay intervals occurring in interactive computing systems.................................................................70

CHAPTER 2: General Method

Figure 2-1 Between-phase (left) and within-phase (right) alternations of experimental conditions ..................................................................................................................83

Figure 2-2 Hypothetical data provided by one subject across four days, using an alternating conditions design presenting two conditions (No Delay and Delay) per day .................................................................103

CHAPTER 3: Effects of Enforced Delays

Table 3-1 Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Percentage of Frames Correct than the Other Two Conditions.........................................................110

Table 3-2 Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Mean Time per Frame than the Other Two Conditions.........................................................115
Table 3-3  Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Satisfaction than the Other Two Conditions .................................................................118

Table 3-4  Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Percentage of Frames Correct than the Other Two Conditions .................................................................126

Table 3-5  Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Mean Time per Frame than the Other Two Conditions .................................................................129

Table 3-6  Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater First-pass Time than the Other Two Conditions .................................................................131

Table 3-7  Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Satisfaction than the Other Two Conditions .................................................................133

Figure 3-1  For each subject, the percentage of frames correct per session (excluding reviewed items), for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Contingent Delay (CD; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison .........................................................109

Figure 3-2  Percentage of first pass frames answered correctly in baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Contingent Delay, CD) for each subject. Median results across subjects are also shown.........110

Figure 3-3  For each subject, mean time per frame per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Contingent Delay (CD; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison .........................................................114

Figure 3-4  Mean time per frame during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Contingent Delay, CD) for each subject. Median results across subjects are also shown. Weighted means have been used because varying set sizes and random assignment of conditions to sets resulted in slightly different numbers of frames in each experimental condition .........................................................115

Figure 3-5  For each subject, satisfaction per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Contingent Delay (CD; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison .........................................................117

Figure 3-6  Mean satisfaction per session during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Contingent Delay, CD) for each subject. Median results across subjects are also shown.................................118
For each subject, the percentage of frames correct per session (excluding reviewed items) for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Noncontingent Delay Blank Screen (NCB; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M”

Figure 3-8

Percentage of first pass frames responded to correctly in baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Noncontingent Delay Blank Screen, NCB) for each subject. Median results across subjects are also shown.

Figure 3-9

For each subject, mean time per frame per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Noncontingent Delay Blank Screen (NCB; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M”

Figure 3-10

Mean time per frame during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Noncontingent Delay Blank Screen, NCB) for each subject. Median results across subjects are also shown. Weighted means have been used because varying set sizes and random assignment of conditions to sets resulted in slightly different numbers of frames in each experimental condition.

Figure 3-11

For each subject, satisfaction per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Noncontingent Delay Blank Screen (NCB; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M”

Figure 3-12

Mean satisfaction per set during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Noncontingent Delay Blank Screen, NCB) for each subject. Median results across subjects are also shown.

CHAPTER 4: Effect of Delays on Workrate and Retention

Table 4-1
Proportion of Sessions in Which Noncontingent Delay had a Greater* First-pass Time, Question Response Time, and Answer Question Time than No Delay

Table 4-2
Proportion of Sessions in Which Noncontingent Delay had a Greater Proportion of Correct Responses (% Correct) than No Delay

Table 4-3
Difference in Accuracy between Noncontingent Delay and No Delay for Experiments 1 to 4.
Table 4-4  Proportion of Sessions in Which Noncontingent Delay had a Greater First-pass Time, Question Response Time, Answer Score Time, and Score Question Time than No Delay ........................................173

Figure 4-1  For each subject, percentage of frames correct per session (excluding reviewed items), for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles). During AC, conditions have been plotted in to facilitate comparison ........................................143

Figure 4-2  Percentage of frames correct in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject ........................................145

Figure 4-3  For each subject, the mean time per frame in each session for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles) ........................................146

Figure 4-4  Mean time per frame in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject ........................................147

Figure 4-5  For each subject, satisfaction per session for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles) ........................................150

Figure 4-6  Mean satisfaction ratings in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject ........................................151

Figure 4-7  For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The same 50-item test was used on each testing occasion.) ........................................154

Figure 4-8  For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). (Baseline, No Delay, Noncontingent Delay, and reversal were represented by 14, 15, 12, and 9 items, respectively) ........................................155

Figure 4-9  For each subject, the percentage of frames correct per session, for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison ........................................166

Figure 4-10  Percentage of frames correct in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject ........................................168

Figure 4-11  For each subject, the mean time per frame in each session for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles) ........................................171

Figure 4-12  Mean time per frame in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject ........................................172

Figure 4-13  Mean satisfaction ratings in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject ........................................175
Figure 4-14 For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The posttest and follow-up test were identical and 60 items long, and the pretest used half these items.) .................................................................175

Figure 4-15 For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). A missing bar indicates a score of zero. (Baseline, Review, No Review, and reversal were represented by 15 items each in the posttest and follow-up test, and by 7, 8, 8, and 7 items, respectively, in the pretest.) .............176

Figure 4-16 For each subject in each experiment, the mean accuracy across the alternating conditions phase for Noncontingent Delay and No Delay ....178

CHAPTER 5: Effects of Reviews

Table 5-1 Proportion of Sessions in Which Review had a Greater First-pass Time, Question Response Time, Answer Score Time, and Score Question Time than No Review .................................................191

Table 5-2 Proportion of Sessions in Which Review had a Greater Proportion of Correct Responses (% Correct) than No Review ..................202

Table 5-3 Proportion of Sessions in Which Review had a Greater First-pass Time, Question Response Time, Answer Score Time, and Score Question Time than No Review .........................................206

Figure 5-1 For each subject, the percentage of frames correct per session, for Review (REV; unfilled circles) and No Review (NR; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M” .....................186

Figure 5-2 Percentage of frames correct in baseline (Base), each experimental condition (Review, REV; and No Review, NR), and reversal (Base2), for each subject .................................................................187

Figure 5-3 For each subject, the mean time per frame in each session for Review (REV; unfilled circles) and No Review (NR; filled circles). Sessions with missing data are marked “M” ......................................................189

Figure 5-4 Mean time per frame in baseline (Base), each experimental condition (Review, REV; and No Review, NR), and reversal (Base2), for each subject .................................................................190

Figure 5-5 Mean satisfaction ratings in baseline (Base), each experimental condition (Review, REV; and No Review, NR), and reversal (Base2), for each subject .................................................................192

Figure 5-6 For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The posttest and follow-up test were identical and 60 items long, and the pretest used half these items.) .................................................................193
Figure 5-7 For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). A missing bar indicates a score of zero. (Baseline, Review, No Review, and reversal were represented by 15 items each in the posttest and follow-up test, and by 7, 8, 8, and 7 items, respectively, in the pretest.) ..............194

Figure 5-8 For each subject, the percentage of frames correct per session, for No Review (NR; unfilled circles) and Review (REV; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison.................................................................201

Figure 5-9 Percentage of frames correct in baseline (Base), each experimental condition (No Review, NR; and Review, REV), and reversal (Base2), for each subject........................................................................................................202

Figure 5-10 For each subject, the mean time per frame in each session for No Review (NR; unfilled circles) and Review (REV; filled circles)........204

Figure 5-11 Mean time per frame in baseline (Base), each experimental condition (No Review, NR; and Review, REV), and reversal (Base2), for each subject.................................................................205

Figure 5-12 Mean satisfaction ratings in baseline (Base), each experimental condition (No Review, NR; and Review, REV), and reversal (Base2), for each subject.................................................................207

Figure 5-13 For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The posttest and follow-up test were identical and 60 items long, and the pretest used half these items.) .................................................................209

Figure 5-14 For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). (Baseline, Review, No Review, and reversal were represented by 15 items each in the posttest and follow-up test, and by 7, 8, 8, and 7 items, respectively, in the pretest.) .................................................................210
Acknowledgments

I sincerely thank:

Dr. John Crosbie for introducing me to behaviour analysis, for modelling scientific rigour, and for providing excellent supervision throughout the project;

Dr. Patrick Flanagan for adopting the role of surrogate supervisor in the latter stages of the project: his patience, friendship, and expert advice during the preparation of this thesis were boundless;

Suzanne Brown for keeping me aware of the world outside the thesis. Her friendship, advice, and consideration were priceless; and

Dr. Simon Parker for being an endless source of tangible support and practical advice.
ABSTRACT

A cornerstone of much educational research in individualised and automated instruction (e.g., computer-based learning) is the assumption that learners be permitted to set the pace at which they work through the material to be learned. Experiments that have compared learning under conditions of self pacing (determined by the learner) and external pacing (determined by the experimenter), using a variety of tasks and populations, often have not supported this assumption. To evaluate the putative advantages of student self pacing in automated instruction, the studies in this thesis compared the effects of self-paced, and externally-paced, programmed instruction on student accuracy, retention, efficiency, and satisfaction.

Under self-pacing conditions, learners completely controlled the rate of progress through learning materials; that is, although the program paused when learners were required to provide answers, score answers, and proceed to the next item, it continued as soon as the learner pressed any key. External pacing was operationalised by programming a noncontingent 10-s postfeedback delay after every item; that is, learners could not progress to a subsequent item until the delay period was over. All relevant learning material for the current item was present during the delay. In a series of experiments using an alternating conditions design, learners completed approximately 40 sets of a programmed course in behaviour analysis (Holland & Skinner, 1961). A baseline of self-pacing conditions was followed by an experimental phase in which baseline conditions were randomly alternated with one or more experimental conditions. Later experiments also included a return to baseline conditions.

In Experiments 1 and 2 externally-imposed delays resulted in greater accuracy than self pacing. This advantage occurred when the delays were accompanied by the study materials, but did not occur for a condition in which delays were presented without the learning material being visible. Hence, it was proposed that noncontingent postfeedback delays are effective because they provide a study opportunity which is otherwise not taken. In addition, imposing delays only slightly increased overall time to completion, and learners rated their satisfaction with external and self pacing similarly.

Experiments 3 and 4 replicated the accuracy advantage found for external pacing, and showed also that material learned under these conditions was recalled better in both immediate and 1-month delayed posttests. These experiments also provided information about factors that influence efficiency during completion of materials. One of these factors was a requirement that, at the end of an instructional
set, each question answered incorrectly be repeated until it was answered correctly (i.e., review feature). This is part of the standard implementation of programmed materials and had been employed in all conditions. In the earlier studies, externally-paced and self-paced conditions showed little difference in overall time to completion. It was apparent that although the externally-paced condition had an increased task time due to enforced delays, this condition did not take longer overall because more errors were made in self pacing, so more items were reviewed, and the overall time of a session was increased. Therefore, although imposing delays entailed a time cost, this was offset because it reduced the number of errors and time-consuming repeats. Experiment 4 demonstrated that when the review requirement was removed, noncontingent delays caused an increase in overall time to completion. Another factor determining efficiency was workrate during nondelay components of the task. Measures of the time learners spent responding, correcting responses, and continuing to subsequent frames, indicated that delays promoted faster workrates at each of these points. This was interpreted as evidence of a generalised escape motivation that is increased by being delayed and which offsets some of the time lost due to delays.

The final two experiments investigated the effects of reviewing incorrect items on student performance because it had been a potential confound in previous experiments. Previously, both self-pacing and external-pacing conditions required subjects to repeat incorrect items until answered correctly. It is possible that because reviewing items increased time on task (like imposed delays), they also led to compensatory changes in workrate, and influenced timing and efficiency measures. Any such influence was not controlled across experimental conditions, however, because self pacing typically resulted in more errors and larger reviews, and any influence of review size on timing measures could not be separated from the effect of delays. It was found that, compared to a no-review condition, reviews reduced efficiency and had little influence on accuracy and retention. Hence, this feature was unlikely to have interacted with the delay variable in previous experiments.

In conclusion, the results of the experiments show that self pacing reduced accuracy, retention, and workrates compared to external pacing. These studies indicate that learners often make poor choices about optimum learning conditions. They also show that small changes in the learning environment can result in consistent and substantial changes in learner performance, and that behaviour analysts have an important role to play in the design and implementation of instructional materials.
CHAPTER 1: LITERATURE REVIEW

Overview

This thesis presents a behavioural investigation of external pacing in individualised instruction. The experiments do not compare instructional techniques or assist the comparison of theoretical positions, but instead assess the effects on performance of self pacing and external pacing during self instruction. A brief history of automated instruction and computer-based instruction is provided along with a consideration of some problems associated with assessing the effectiveness of instructional software. Following this, it is demonstrated that there is a great deal of dissatisfaction with the quality of instructional software among its users, and it is proposed that, at least in certain types of software, dissatisfaction could be avoided by incorporating well-established principles of instruction. The thesis then focuses on the prevailing notion that instruction should be completely self paced; this is one of the fundamental tenets of behaviourally-prescribed approaches to instruction and is the prescription that continues to be adopted by most current software applications. Evidence is provided that self pacing has not lead to optimal performance in a variety of verbally-based tasks, instead, external control of pacing by imposing delays is beneficial. Hence, the self-pacing-is-mandatory assumption in self-instructional materials is questioned. An experimental assessment of the effectiveness and efficiency of externally controlling student pace by imposing delays is then presented.

Although the thesis is primarily concerned with the pacing issue in instruction, it was necessary to make two preliminary decisions about the materials and medium to be used. Programmed instruction is the instructional approach that will be employed because it has a strong theoretical and empirical basis, and because the notion that students should determine their own study pace has been derived largely from this theoretical base. It is logical to assess the notion in the context in which it originated. Teaching machines were designed to present programmed materials, but computers will be used as the vehicle of instruction in this thesis because they can both replicate the functions of teaching machines, and also offer many presentation and recording benefits (which are discussed later).

Automated Instruction

Although teaching aids have been used for centuries (Buck, 1990), the devices developed by Pressey and Skinner within this century are most relevant to the present trend towards computer assistance in teaching. Each of these devices
were based on established psychological principles, were well researched, and were aimed at improving performance and efficiency in the classroom.

Pressey's Contribution to Automated Instruction

In the mid 1920's, Sidney Pressey wrote that he would like to see "fine enthusiasms, clear thinking and high ideals" (1926, p. 376) developed in students. He considered that it was the teacher's task to nurture these qualities, but proposed that this would only be possible if they could be freed from the time-consuming burden of teaching drill material and dealing with paper work. To integrate teaching with time- and labour-saving machinery, Pressey developed a device which he claimed "gives tests and scores and teaches" (1926, p. 373). This device was exhibited at the 1924 meeting of the American Psychological Association. In the following year a modified version of the device was presented and a paper describing its operation was subsequently published. A schematic view of this device is shown in Figure 1-1.

Figure 1-1. Pressey's automated teaching device.

Pressey's device was about the size of a portable typewriter and looked like two cylindrical casings: the first slightly wider than a sheet of paper with a viewing window; this cylinder was connected to a second, shorter casing from which protruded four response keys labelled "A", "B", "C", and "D". A drum inside the longer casing was enveloped by a sheet of paper with thirty questions and multiple-choice answers. An example of a question is, "To help the poor debtors of England James Oglethorpe founded the colony of (1) Connecticut, (2) Delaware, (3) Maryland, (4) Georgia." (Pressey, 1926, p. 374). One question and the alternative answers could be seen through the window. Students pressed one response key to answer. If the answer was correct, then the drum inside the casing rotated the next question into view. If an incorrect answer was selected, the question remained visible until a correct response was made. Hence, Pressey's device employed a respond-until-correct procedure. The question set was presented in full, at least twice. After two passes through the items, any question which had not been answered correctly the first time, on both occasions, would be presented repeatedly until criterion was met. This criterion could be set higher if required. After students completed the study material, the device issued a coupon to indicate mastery. More
frequent rewards could also be issued by setting the machine to deliver small pieces of candy for correct responses. If the device was used to administer a test, only one response per item was possible, and the next question was then presented.

According to Pressey (1927) his apparatus was “built as to function, . . . in accordance with what is now known concerning the learning process” (p. 551). Although it is unclear precisely what Pressey’s theoretical support was, his writings (1926, 1927) strongly indicate the influence of Thorndike’s laws of learning. Pressey maintained that, because students were penalised for incorrect answers by having to repeat the question, but were rewarded for meeting the criterion for correct responses by elimination of the question, the law of effect operated. Because students had to give the correct response once on each trial in order to progress, it was the response made most frequently, and, hence, the law of exercise also was utilised. Pressey also wrote that the law of recency was followed because the most recent response to an item was necessarily the correct one and this established “the correct answer in the mind of the subject” (1926, p. 375). The device also supported three components of efficient learning: (a) the student could progress at a pace determined by their own needs, rather than the pace of the class; (b) immediate feedback about the correctness of responses (cf. waiting for return of a teacher-corrected paper); and (c) removal of a question from presentation once a mastery criterion had been met. This eliminated what Pressey considered wasteful overlearning, and learning could be individualised.

Pressey’s machines made two important contributions to education: (a) they provided technical support for teaching by presenting and correcting papers; and (b) they showed that machines could implement effective instructional principles. Pressey was optimistic about the future of labour-saving devices in education and the emancipation of teachers from routine tasks. He envisioned mechanical aids for students of arithmetic, spelling, language, and for the collection of data (1932). This enthusiasm, however, was short lived. His devices held the attention of the research community only briefly (Benjamin, 1988; Holland 1960), and, in 1932, Pressey publicly voiced his disappointment at society’s failure to embrace a potential “industrial revolution in education”:

The problems of invention are relatively simple; with a little money and engineering resource, a great deal could easily be done. The writer has found from bitter experience that one person alone can accomplish relatively little, and he is regretfully dropping further work on these problems. But he hopes that enough may have been done to stimulate other workers, that this fascinating field may be rapidly developed. (p. 672)
According to Skinner (1958) the general acceptance of Pressey's machines was hindered by the cultural and psychological climates of the time. First, educators and researchers were not ready for Pressey's advanced concept, and use of the machines declined because of "cultural inertia". One reason for this inertia was that Pressey was advocating classroom efficiency (faster education with a need for fewer teachers) at a time when the country was in the Great Depression, and there was a surplus of teachers and no pressure to increase the pace of education (Benjamin, 1988). Second, Pressey's machines were limited in function. They were developed when rates of learning and forgetting were of interest and reflected the dominant learning theories of the time (e.g., laws of recency and frequency). As a result the devices were versions of a memory drum and were appropriate for studying the rate of learning, and testing previously learned material (Skinner, 1958). In the decades following Pressey's work, a psychological understanding of learning processes developed, primarily from the work of Skinner, and it became clear that Pressey's machines had limited potential for teaching new material. Some years later, Holland (1960) alluded to a third limitation of early attempts to automate teaching: too much emphasis was placed on the gadget and too little on the technique (of behavioural control over the student). Perhaps it was too easy for Pressey's intended audience to associate gadget with fad and underemphasise the educational contribution of his machine.

Although the technical contribution of Pressey's machine was important (by removing rudimentary and routine tasks from teachers' duties), its pedagogical potential was greater. It was demonstrated that a machine could allow for individual pacing, ensure active participation from students, provide immediate feedback in the learning process, and test for mastery of one section of material before progressing to the next. Some of these are still central aspects in prominent, behaviourally-based teaching methods (e.g., Personalized Systems of Instruction, PSI; Keller, 1968) and some varietics of Computer-Based Instruction (CBI). It was to be more than 20 years, however, before the next major advance in automated teaching occurred.

**B. F. Skinner's Contribution to Automated Instruction**

When Skinner developed his teaching machines he was unaware of the earlier work of Pressey; his entry into the field was personally motivated. In 1953 Skinner attended his daughter's fourth-grade arithmetic class and, after observing the students at work, recognised two problems which he considered detrimental to the progress of students: they were expected to progress at the same pace (which was too slow for some and too fast for others), and they had to wait until the next day to
receive their corrected papers and knowledge about their performance (a detailed account of this experience is given in Skinner, 1983). Not only was feedback delayed but there was a prevailing lack of it. For instance, when solving a mathematics problem, feedback received at the point of solution may not be relevant to the earlier steps—all of which need to be successful for a correct final solution. Skinner considered feedback of response correctness analogous to a reinforcer (that is, a stimulus that increases the likelihood of the behaviour it is contingent upon), and, under good learning conditions, complex behaviour is efficiently shaped and promoted by reinforcing each of its component behaviours. With complex verbal behaviour in the classroom, feedback needed to be contingent on many more of the steps leading to the final result. Because it was beyond the capacity of a single teacher to provide the amount of reinforcement required to build complex behaviour (Skinner, 1954, 1961), Skinner’s self-assigned task was to determine how to reinforce correct behaviour at the time at which it is performed, for each of the many students in a class. His solution was twofold: arrange the materials to be learnt in an orderly and economic manner, and establish a device that can deliver the materials according to effective contingencies of reinforcement.

Programmed Materials

Skinner proposed that there are several requirements for the efficient use of teaching materials: (a) divide the subject matter into small units, include only essential facts, terms, and laws, with some examples, and remove all superfluous material; (b) arrange these units in order of complexity so that a student moves step by step from ignorance to knowledge; (c) present the units in frames (i.e., a small amount of learning material, one question and answer), so that students must continually be actively responding; (d) include immediate feedback for all appropriate responses; and (e) allow students to move through these materials at their own pace. Materials organised in this manner became known as programmed materials, as they allowed the student to progress gradually with a minimum of errors (Holland, 1963).

Implementing Reinforcing Contingencies

Construction of materials in this way was not enough to ensure learning, however. Presentation of materials and consequences of student behaviour needed to be governed by appropriate contingencies of reinforcement. To do this, Skinner adapted several general and reliable principles of learning that had been established in the operant laboratory to verbal behaviours.

According to Skinner, the theories of learning represented in Pressey’s machine were outmoded. During the years separating the release of Pressey’s and
Skinner's machines, Skinner and others had established various techniques that enabled the precise control of the operant behaviour (i.e., behaviour that operates upon the environment to produce reinforcement) of both human and nonhuman subjects. Shaping is one such operant technique that entails progressively reinforcing responses in the direction of the response pattern required. Although Skinner's techniques (e.g., reinforcement, shaping) were often associated with nonhumans and nonverbal behaviour, it can be argued that they also were appropriate for education. Teaching is the modification of education-relevant behaviours and, although this behaviour is often verbal, verbal repertoires are controlled by the same techniques (such as reinforcement and shaping) as nonverbal behaviour (Holland, 1960; Skinner, 1957, 1989). By this reasoning Skinner's techniques were appropriate for education.

The following principles were the result of Skinner's application of his operant research to his teaching technology (Holland, 1960; Skinner, 1958):

1. **Immediate reinforcement**: for a response to be learned, it must be reinforced immediately so that there is a contingent relationship between the two events. Reinforcement, at high rates, for correct performance maintains responding.

2. **Active participation**: learning requires the reinforcement of a behaviour, and reinforcement can be delivered by an external source only when the behaviour is overt. Traditional educational media (e.g., lectures, textbooks) do not produce overt behaviours and may continue to present information even when students are inattentive, whereas a machine necessitates overt responses and reinforces them when correct.

3. **Gradual progression**: complex operant repertoires (i.e., complex behaviour) can be systematically established by the ordered presentation of materials from the simple to complex in small manageable units.

4. **Prompting/fading**: students' reliance on external help when responding is removed by gradual withdrawal of stimulus support (i.e., "vanish", "fade", or remove from the stimuli, hints and clues to the correct response).

5. **Control/attention**: programs and machines control students' observing and echoic behaviour. They must observe or "attend" the information in front of them and respond with respect to it or they will not progress. This is contrary to lectures, for example, which progress independent of student behaviour (e.g., they may "complete" a lecture even though they sleep through it).

6. **Discrimination training**: programmed instruction involves discrimination training: Learners examine a stimulus, emit a response, receive feedback, and
consequently make appropriate discriminations in the presence of novel stimuli. This is assisted in programmed materials by the use of many examples and discussion of concepts from different perspectives so that responding is not controlled by only a limited stimulus set.

The use of programmed materials according to these principles became known as programmed instruction. In addition, two educational doctrines accompanied these principles (Skinner, 1968):

1. *Self-pacing*: because students differ in their rates of progression, they should be allowed to work at their own pace. In this way, quicker students are not restrained and do not become bored as a result of a slower class pace, and slower students are not left behind by a class pace which they cannot manage.

2. *Nonaversive*: aversive control is prevalent in schools (in older days the birch rod and cane prevailed; more recently ridicule and teacher displeasure have been the major aversive methods) but does not facilitate learning, hence, only positive reinforcement should be used.

**A Typical Skinnerian Teaching Machine**

To present both the programmed materials and the contingencies of reinforcement, Skinner developed a teaching machine. Although there were numerous variations on Skinner’s teaching machine (e.g., see Holland, 1960; Skinner, 1958), all utilised the principles outlined above. In one version (see Figure 1-2), the machine appeared vaguely like a typewriter but the keys were replaced by a flat surface with one window exposing the learning material and a question. An adjacent window exposed a paper strip on which a response could be written. When a sliding knob above the response window was moved the correct answer was presented next to the student’s response, and a transparent sheet covered the existing response to avoid alteration. Students could mark the correctness of their own response in the response window. A dial towards the top of the machine scrolled the next frame into the viewing window. After a student had completed all frames, those that were answered incorrectly were presented again. This continued until all frames were answered correctly. Frames and responses were written on continuous paper and, once a session had begun, the machine could not be opened without breaking the answer strip. This record of student responses and correct answers was a permanent record of performance which could be checked by a course supervisor. According to Skinner (1961), this teaching machine was unusually efficient because it provided frequent and immediate reinforcement, allowed students to move at their own, natural pace, and follow a coherent sequence of information.
Like Pressey, Skinner recognised the need to increase the efficiency of current teaching practices:

There are more people in the world than ever before, and a far greater part of them want an education. The demand cannot be met simply by building more schools and training more teachers. Education must become more efficient. To this end curricula must be revised and simplified, and text-books and classroom techniques improved. In any other field a demand for increased production would have led at once to the invention of laborsaving capital equipment. (Skinner, 1958, p. 29)

The differences in the approaches to instruction of Pressey and Skinner were fundamental, however, and went beyond instrumentation. Programmed instruction used many small steps to enable students to have a high success rate and maintain motivation; Pressey on the other hand was not overly concerned with students making errors because, he argued, they would eventually find the correct answer (he also considered the feature of allowing multiple errors on one question useful for tests; 1926). The way responding was operationalised in the machines also was a major difference. Skinner objected to Pressey's use of multiple-choice responding because it exposes students to options which are plausible but wrong--something that is confusing and that increases the likelihood of errors ("Ice Cream," 1968; Skinner, 1961). He advocated constructed responses (writing the answer without the aid of options) because this skill has more utility in life than pressing buttons (multiple choice). Programmed instruction was designed to take someone who knew nothing about a subject and precisely control and develop their behaviour until they had a complex response repertoire. Pressey’s machines, however, were predominantly labour-saving devices, they provided quizzes to students who had previously learned the material elsewhere. The use of machines for providing instruction rather than saving labour after learning had occurred was fundamental to Skinner: “The underlying principle of programmed instruction, which Pressey
rejected, seemed to me more important than the machine" (Skinner, cited in Epstein & Skinner, 1982, p. 222).

Both Pressey and Skinner recognised the need for individualised instruction, a need that is gaining increased attention today as education budgets decrease, literacy problems escalate, and computing technology diminishes the distinction between distant and on-campus learners. Other common elements in their approaches were active responding, immediate feedback, and self pacing. Self pacing assumes (a) that the class rate does not suit all students, which is probably true, and (b) that students can determine an appropriate pace for learning. The self-paced tenet will be fully addressed later. Next, the impact of programmed instruction is considered further.

The Influence of Skinner's Contribution

Skinner’s work made a large contribution to individualised instruction. It is important to note that, beyond the technology itself, there were several factors that probably aided societal acceptance of Skinner’s approach which were not present in Pressey’s time: widespread establishment of audiovisual aids in schools had paved the way for teaching machines (Benjamin, 1988), Ramo (1957) had published an influential paper advocating technology and machinery in schools, and there was a critical shortage of teachers (Kreig, 1961).

There are a number of dimensions on which Skinner’s contribution to individualised instruction can be judged. Research generated is one example. Prior to 1948 there had been only six published papers on teaching machines, and fewer than eight papers per annum were reported in each of the subsequent 10 years (Fry, Bryan, & Rigney, 1960). But research blossomed, and by the mid 1960’s programs had become available for every subject at each grade level, and more than 300 papers a year were being published (J. A. Kulik, Cohen, & Ebeling, 1980). In addition to the scientific literature there were many more reports each year about Skinner’s technology appearing in popular magazines (Benjamin, 1988).

Skinner’s work also stimulated interest among publishers and manufacturers who saw that his device had commercial possibilities. By 1962 there were nearly 200 private companies producing programmed materials, teaching machines, or both (Benjamin, 1988; see also Smith & Moore, 1962, Appendix B). The market was offering dozens of electronic and mechanical teaching machines ranging in price from $20 to $6500 (Kreig, 1961). One major American encyclopedia publisher hoped to sell $5 million in teaching machines and programs in their first year ("The teaching machines," 1960). Complex commercial models were
constructed by large companies such as International Business Machines (IBM), Rheem Company, and Bell Telephone Laboratories. A bibliography of commercially developed teaching devices can be found in Fry et al. (1960). So, in terms of the amount of research generated and applications of the principles of learning to automated devices, programmed instruction did have a large impact. But while commercial and research outcomes are important, the most important measure of Skinner’s contribution to education is the instructional effectiveness of programmed instruction.

**Instructional Effectiveness of Programmed Instruction**

There is considerable anecdotal and experimental evidence regarding the effectiveness of programmed instruction.

**Evaluation of Applied Uses of Programmed Instruction**

Programmed instruction was first assessed in 1958 with Skinner’s Natural Sciences 114 course at Harvard. Previously, a conventional textbook (Science and Human Behavior: Skinner, 1953) had been used but this was converted to program form with the assistance of James Holland (Holland, 1976). Appropriately, the program content was the experimental analysis of behavior, and Harvard and Radcliffe undergraduates worked with the program to prepare them for a better understanding of the lecture portion of the course.

At the beginning of the semester students were informed of the program that was to be completed in the two months preceding lectures (Skinner, 1983). The program was divided into many discrete sets, each covering one main topic and requiring about 20 mins to complete. Completing all sets required a median time of 14.5 hours (Skinner, 1958). Students could complete the sets at their own pace on teaching machines in the self-instruction room at Harvard. This room was operated by a course assistant who supplied students with program disks to load into machines, and collected tape receipts containing their written answers. Although students could complete sets at their own pace, the sequence was predetermined, and mastery of each set was a prerequisite for progression to the next.

Skinner reported that, as a result of program use, students understood the principles about which he lectured. Furthermore, interview and questionnaire data from these students revealed that they gained much by machine study. Compared with conventional methods, they claimed to learn more in less time and with less effort. Interest was maintained without special incentive, and the continual knowledge of progress was much appreciated (Skinner, 1958). Additionally, this method provided item-by-item records of each student’s responses, and gave the authors a precise means by which they could improve the program for future
courses (Holland, 1960; Holland & Doran, 1973). As a result, Skinner continued to use the programmed course in future years.

Skinner was convinced that programmed instruction was better than traditional instruction, and that students could learn twice as much in the same time and with the same effort used in the traditional classroom (1961, 1966). "The Roanoke experiment" is the most often cited test of Skinner's conviction (Kreig, 1961; Rushton, 1965; Skinner, 1961, 1983, 1984, 1989). The Roanoke experiment was a field trial of programmed instruction conducted in 1960 in Roanoke, Virginia. Students in an 8th-grade class used primitive teaching machines and hastily-prepared materials to learn 9th-grade algebra. Working at their own pace, the students completed one year of 9th-grade algebra in one term. They scored at least average on the 9th-grade norms. One year later, those who were retested had a retention rate of more than 90% (where normal was 70%-80%; Skinner, 1983). Because impressive results from programmed instruction continued in subsequent years, improved performance was unlikely to have been a result of novelty. Others have documented similarly impressive findings regarding student interest and long-term outcomes using programmed instruction (e.g., Hawkridge, 1966, cited in Schramm, 1977).

The Harvard and Roanoke accounts describe pioneering attempts to automate and individualise instruction in the classroom. They indicate that both staff and students benefited from using programmed materials as a substitute for traditional instruction. There also were many experimental comparisons of programmed and traditional instruction, and these are discussed next.

Experimental Evaluations of Programmed Instruction

During the programmed-instruction movement, many studies attempted to determine experimentally whether this method taught students better than conventional methods. These studies were examples of what are often referred to as media-comparison studies where one instructional medium (e.g., teaching machine, computer) is compared with another (typically a traditional classroom). In a typical comparative study evaluating programmed instruction, a teacher divided a class into two groups. One group would be taught by traditional instructional methods, and the other would work from programmed materials (generally a textbook version). After the relevant material had been covered by all students, they would take a common exam so that their relative achievement levels could be compared. Other measures often included immediate learning, time taken to complete materials, and students' subjective evaluation of the method. Outcomes from these studies typically showed that programmed instruction resulted in performance equal to, or
better than, traditional methods. It also was found that programmed instruction was more time efficient, and students often liked the method. Table 1-1 summarises the achievement outcomes from several reviews of comparative studies. A discussion of the review findings follows.

Silberman (1962) reported outcomes from 15 comparative studies: In 9 studies, the programmed-instruction group showed greater learning, there were no studies in which a traditional group performed best, and 6 studies found no differences between groups. Programmed instruction required less time in all studies. Similar findings were reported by Schramm's (1964) review of 36 studies: 17 reported significantly greater achievement due to programmed instruction, 1 reported significantly better performance under traditional instruction, and 18 indicated no significant differences between groups. Eight experimenters reported shorter completion times for programmed instruction.

In a more comprehensive survey, Nash, Muczyk, and Vettori (1971) summarised 213 comparisons (over 150 studies) of traditional instruction and programmed instruction. Of the 138 comparisons of immediate learning, 49 found a significant difference in favour of programmed instruction, only 18 favoured traditional instruction, and the most common result, found in 71 comparisons, was that there was no significant difference between the two methods. A similar result arose when retention outcomes were compared. Most studies (90%) found a training time advantage for programmed instruction. Unlike previous reports, this review considered not only the direction of a difference between methods, but the magnitude of this difference. It was found that learning advantages ranged from a 20% difference favouring traditional methods to a difference of more than 50% favouring a programmed approach. Such differences could be explained, at least in part, by experimental features such as setting (academic, industrial), trainee (e.g., high and low ability), implementation (e.g., whether programmed instruction was a supplement to, or a substitute for, the traditional method), and control procedures used in the studies. Although the findings provided some evidence of time and performance advantages for programmed instruction, Nash et al. concluded that "the most evident conclusion about the relative effectiveness of programmed instruction is that research completed on the topic is methodologically very poor" (p. 408). They highlighted problems such as overly small sample sizes, lack of control groups, and lack of control over important variables (such as time spent on materials), as contributing to a limited interpretability of the studies.

The previous evaluations of programmed instruction indicated that, although it often led to better performance than traditional instruction, often it did not. A
satisfactory generalisation about the effectiveness of programmed instruction could not be made, and a potential consumer could not predict whether, or to what extent, this method would benefit them. Although Nash et al. indicated that poor research contributed to variability in findings, it is also true that poor evaluative tools were partly responsible for the difficulty in describing the general effects of programmed instruction, and the variables that influenced its effectiveness. The following section describes a different evaluative tool designed to overcome the shortcomings associated with reviews such as those described above.

**Meta-analyses vs. frequency counts.** There are two ways to evaluate comparative studies. Reviews like those discussed above determined which was the more effective technique using what is known as a frequency count (or voting, tally, or box score, method). That is, for each relevant dependent variable (e.g., exam achievement), it was determined whether programmed instruction or traditional instruction scored higher, or, whether there was no significant difference between the two. The final counts should indicate whether one method is reliably superior to the other.

This sort of review has three problems, however: First, nonexhaustive reviews can result in a nonrandom sample of studies, so objective conclusions regarding the relative merits of each technique is impossible. Second, the frequency count does not indicate whether any differences in the success of the techniques are substantial or practically relevant (although Nash et al., 1971, made attempts to overcome this). Finally, frequency counts offer no objective, statistical means of determining which characteristics distinguish effective from ineffective use of programmed instruction. This form of analysis assumed that any influential factors (such as quality of a program or teacher) would, in a large sample, cancel each other out and leave a general trend attributable to the instructional method.

A more recent analytical method of reviewing large numbers of studies is known as meta-analysis (Glass, 1976). This is a statistical technique that is designed to overcome limitations in experimental reviews like those discussed above. Meta-analysis involves the statistical aggregation of findings from all relevant studies so that an overall conclusion can be reached. To ensure the reliability of studies used in a meta-analysis, they are located and screened for inclusion using clearly-specified methods and criteria. All potentially influential characteristics of the studies (e.g., methodology, ability of subjects, treatment duration) are then coded so that they can be related statistically to the summary outcome, and each outcome variable (e.g., exam achievement, satisfaction) is expressed in quantitative terms (e.g., achievement score of a class using
programmed instruction minus achievement score of a class using traditional instruction). A summary statistic condensing the information from each study can then be calculated. One such statistic is called the effect size (ES). Although there are variations of the statistic, a common way to calculate an ES is by dividing the difference between the performance scores of an experimental and control group by the standard deviation of the control group (e.g., C-L. C. Kulik & Kulik, 1991; see Figure 1-3). In other words, the ES indicates the number of standard deviations, on average, by which a treatment score differs from a standardised mean. This z-score unit means that effects from different studies using different dependent variables can be compared.

\[
ES = \frac{\bar{X}_e - \bar{X}_c}{s_c}
\]

**Figure 1-3.** Formula for calculating effect size.

Some years after the programmed instruction movement had ceased, Hartley (1977) used meta-analysis to condense the findings of 40 studies published between 1962 and 1974. These studies investigated the relative effectiveness of programmed instruction and traditional instruction in mathematics education as measured by posttest performance at the end of a course. Using the frequency count method with the information available in the dissertation it is apparent that 22 studies favoured programmed instruction and 18 studies favoured traditional instruction. In addition, a meta-analysis of the outcome findings showed that, on average, students taught with programmed materials performed 0.11 standard deviations higher on examinations; this equates to only about four percentile points, a negligible amount (J. A. Kulik, Kulik, & Cohen, 1980).

Following Hartley’s review, J. A. Kulik, Cohen, & Ebeling (1980) conducted a comprehensive analysis of the effectiveness of programmed instruction on college-level learning. From a pool of over 5000 articles in the area, 57 studies met stringent methodological criteria for inclusion in the analysis. Of the 56 studies that measured exam performance, 21 found a significant advantage for programmed instruction, 4 found a significant advantage for traditional methods, and 31 found no significant differences. In terms of the magnitude of effect, over 30% of the studies reviewed showed a medium to large effect of programmed instruction, over 60% found small effects, and about 5% found moderate to large effects for
traditional instruction. The average effect (i.e., ES) of programmed instruction was to increase the performance of students by about 0.24 standard deviations beyond students learning via traditional methods. Convention holds that 0.2, 0.5, and 0.8 represent small, medium, and large differences between means, respectively (see Cohen, 1977, 1988), and the reviewers described the 0.24 value obtained using programmed instruction as "not spectacular" (p. 62).

Conventions are arbitrary and relative to the field of inquiry, however. The ES can be converted to a percentile change by multiplying it by 34 (the distance, in percentage, of one standard deviation above or below the mean). By doing this, the average effect can be seen to equate to a shift from the 50th to the 60th percentile on a standard distribution. This is a shift of one grade point which may be considered educationally significant.

There are two clear outcomes of each of the reviews discussed above (also see Table 1-1): (a) The overwhelming finding was that about half the studies find no differences between methods, about half favour the programmed approach, and very few favour traditional methods; and (b) ES's were small but were positive, showing an advantage for programmed instruction.

Table 1-1
Percentage of Studies in Which the Effect of Programmed Instruction (PI) and Traditional Instruction (TI) on Achievement was Statistically Equivalent, or Statistically Significant in Favour of One Method

<table>
<thead>
<tr>
<th></th>
<th>PI = TI (%)</th>
<th>PI &gt; TI (%)</th>
<th>PI &lt; TI (%)</th>
<th>Effect size&lt;sup&gt;a&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silberman (1962)</td>
<td>40</td>
<td>60</td>
<td>0</td>
<td>-</td>
</tr>
<tr>
<td>Schramm (1964)</td>
<td>50</td>
<td>47</td>
<td>3</td>
<td>-</td>
</tr>
<tr>
<td>Nash et al. (1971)</td>
<td>51</td>
<td>36</td>
<td>13</td>
<td>-</td>
</tr>
<tr>
<td>Hartley (1977)&lt;sup&gt;b&lt;/sup&gt;</td>
<td>-</td>
<td>55</td>
<td>45</td>
<td>.11</td>
</tr>
</tbody>
</table>

<sup>a</sup>Effect sizes for meta-analyses reporting an advantage of programmed instruction over traditional instruction. <sup>b</sup>Hartley reported only the percent of studies favouring traditional instruction, the other percentages are inferred.
Why did programmed instruction work when it did? An important advantage of the meta-analytic technique is that it may indicate what variables contribute to successful implementation. Both Hartley (1977) and J. A. Kulik, Cohen, and Ebeling (1980) found a moderate, positive correlation between year of publication and effect size of programmed instruction (i.e., more recent studies showed a more favourable effect of this technique). Hartley also found that more recent materials were associated with larger effect sizes. These findings may indicate that people were learning to use the technique more effectively. Perhaps, programmed instruction was used indiscriminately when it was new, under the assumption that it was an educational panacea, but after the excitement had died down and more research had been conducted, it was used more appropriately. For example, it was established that programmed materials were more appropriate for teaching unfamiliar or technical material (Tobias, 1973). It is also reasonable to suggest that the quality of programs improved over time; program quality has a demonstrable effect on student outcomes (Kemp & Holland, 1966). Without consideration of these influences, the inconsistent findings reported in comparison studies may lead to conclusions about the (in)effectiveness or (un)reliability of programmed instruction which may be more appropriately directed at the way it had been used. Although programmed instruction typically resulted in similar or greater levels of achievement than traditional instruction, failure to isolate the key variables controlling positive effects may have been a factor that assisted in the eventual downfall of this movement.

Summary of Instructional Effectiveness

The systematic basis of programmed instruction was its most valuable asset. As a behavioural scientist, Skinner proposed that the solution to America's educational woes would be found in analysing educational behaviour and finding the rules which controlled it. Although programmed instruction stimulated a great deal of research and commercial interest, and demonstrated that it could instruct effectively in applied situations, interest in the technique has all but disappeared. Unfortunately, research did not progress far enough to determine what factors reliably contributed to successful implementations of programmed instruction. Although meta-analytic techniques offered some hope of statistically relating experimental variables to outcomes, this was not fruitful, possibly because the number of studies meeting the strict criteria for inclusion was small, or because important components of the technique were not identified or consistently employed.
Today, teaching interest still lies with prominent technology, but this is now in the vein of computers, animation, and videodisc technology. The programmed instruction model maintained some lasting influence as the basis of much computer-based software (Merrill, 1985), but this influence has been limited because the psychological learning principles that Skinner emphasised as the core of this method of teaching are often not employed in software (Tudor & Bostow, 1991; Vargas & Vargas, 1991). Whether investigation of student behaviour will receive a similar emphasis again is unclear. As educational computing moves increasingly towards CD ROM and exploratory learning formats, the use of behavioural approaches in future educational technology is likely to be further reduced. Coupled with this is an apparent decline in interest in educational research. The number of behaviour analytic studies of academic behaviours published in the Journal of Applied Behavior Analysis has declined steeply over the last three decades (Northup, Vollmer, & Serrett, 1993) even though concern over the effectiveness of the educational system is increasing.

Before concluding this section on Skinner's contribution, it is important to consider the decline of programmed instruction. This is interesting because it appears to have been due to many factors, and, importantly, few of them were related to the instructional effectiveness of programmed instruction. The following section considers the decline in commitment to this technique and its legacy for individualised instruction.

The Decline of Programmed Instruction

By the late 1960's the lack of commitment to programmed instruction was becoming evident as the number of scientific and popular publications about the movement dwindled, and many teaching machines disappeared from the market (Benjamin, 1988). By 1980, programmed instruction was no longer a focus in educational discussion, and papers on the subject had become rare (J. A. Kulik, Cohen, & Ebeling, 1980). Although Skinner maintained that he did not know why American education did not latch onto programmed instruction (cited in Zientara, 1984), his extensive writings on the issue indicate that he considered several factors to be responsible. These factors can be broadly classified as political, intellectual, and social.

Political Factors in the Decline of Programmed Instruction

As Skinner and other proponents were promoting programmed instruction, the political tide changed with the appearance of the Russian satellite, Sputnik, in 1957 which informed America that it had been beaten into space. To Americans this was evidence that their educational approach was faulty and needed to be corrected
(Higginbotham-Wheat, 1990). As a consequence, Congress passed the National Defence Education Act and provided funds to improve teaching and develop new materials. The time after Sputnik has been described as a “cognitive movement” (Skinner, 1986) where discovery learning was emphasised over directing students, and creative thinking preferred over memorisation. This shift in instructional approach meant that a systematic behavioural analysis of teaching and learning was hindered because behavioural techniques were losing attention and funding to cognitivist approaches to psychology (Skinner, 1984).

The educational advantages are not apparent, however: Twenty years later, in 1983, the National Commission on Excellence in Education reported that the average achievement of American high school students on standardised tests was lower than it had been in the 1960's. Furthermore, the downturn was in many fields, including science, and American students compared poorly at an international level (Skinner, 1984). It has been argued that the materials were not at fault, but that the method of teaching was (Skinner, 1984, 1986, 1989). Ten years after the National Commission's report the International Assessment of Educational Progress survey placed American 13-year-olds below those of 13 other countries in science proficiency (Schrage, 1994). This educational crisis is not restricted to the U.S. A recent Australian report indicates that at least 25%, and in some schools, a majority of the students finish primary school with literacy problems (“The Literacy Challenge", 1993). In a prior inquiry into workplace literacy, evidence indicated that between 10% and 20% of Australian adults are functionally illiterate (“Words at Work", 1991). Whether existing methods (e.g., whole language approach) for teaching literacy in schools were successful could not be determined because objective means of assessing literacy attainment were not in place. For the same reasons, comparisons between different methods, or across nations, cannot be made. At the same time that large problems with the educational system are apparent, contributors to Journal of Applied Behavior Analysis question why decisions regarding educational practices continue to be based on dogma rather than scientifically established techniques (Geller, 1992, The education crisis.)

**Intellectual Factors in the Decline of Programmed Instruction**

Programmed instruction was also stifled by intellectual misunderstanding of the technique. Although the laws of learning had convinced behavioural educators that student behaviour could be quickly modified and maintained, they had not convinced many psychologists and educators who appeared either misinformed or ignorant of the many years about work from which programmed instruction was devised (see Skinner, 1983, pp. 200-202).
Proponents of programmed instruction damaged the cause by automating courses (i.e., making them program like) without employing the principles of behavioural control that were responsible for their effectiveness (Holland, 1960). Manufacturers also compromised the technique: For example, a machine feature which enabled a student to review incorrect items was often omitted because it was costly to implement. This feature was responsible for a 10% improvement in performance of students using it compared with those who did not (Holland & Porter, 1961).

Social Factors in the Decline of Programmed Instruction

Programmed instruction also raised many social concerns. Proponents of humanistic views protested against the analyse-and-control approach attributed to Skinner's methods because they considered them a threat to personal freedom and dignity (Skinner, 1984). Figure 1-4 shows a tongue-in-cheek portrayal of a feared outcome of a scientific and mechanised approach to teaching. A new way of teaching also raised many issues for teachers and the educational establishment.

![Figure 1-4](image)

Figure 1-4. Teaching machine and programmed instruction cartoon.


Teacher concerns. Two important concerns for teachers were whether an automated and mechanised approach to instruction made them redundant, and whether it was appropriate to have students being educated by machines rather than humans. Fears of being replaced and losing time with students were unfounded. Both Pressey and Skinner advocated automated devices as aids to improve efficiency and minimise clerical work, so that, rather than being replaced, teachers
could spend more time teaching their students. But Skinner did not communicate successfully the difference between good teaching (by way of contingencies of reinforcement) and personal attention (1980, pp. 97, 273). Had he (and the programmed instruction movement) been more successful at emphasising the importance of the teacher’s role, it could have eliminated some of the antagonism towards it (Leib et al., 1967). Reports from teachers actually using programmed instruction showed that they were happy, in charge, in use, and in no danger of replacement (Kreig, 1961; Rushton, 1965).

Teacher resistance to technology continues to be an important issue. Ellis (1992) reports that many teachers actively resist the use of computers in education, probably because they are computer phobic or are threatened by loss of employment.

Concerns for the educational establishment. A more pertinent problem than dealing with displaced teachers, was deciding what to do with students. If they did learn twice as fast, how long would they be kept in school? What else would they be taught to fill in their years? How soon should they enter the workforce? Furthermore, if all students operate at a different pace, how should they be grouped? The class level, which is intended to describe the extent of student learning achievement, would be redundant. These seem to be more enviable problems than those resulting from widespread, poor student achievement (Skinner, 1984, 1986).

School authorities were in an awkward position. If programmed instruction worked well, then some of the difficulties considered above may become a reality, but if programmed instruction turned out to be a poor innovation, any administrative staff responsible for the project would carry the student’s plight and the guilt of the decision (Skinner, 1983, pp. 167-168). They also were aware that whoever controlled programming also controlled education—was it to be schools, parents, industry, or otherwise (Kreig, 1961)? School administrators also had financial restrictions which meant that innovations that were not guaranteed effective would not receive funds.

Perhaps the biggest obstacle any new teaching technique must overcome is the inertia and rigidity of educational institutions. Overcoming this inertia is extremely difficult according to Keller (1982) who faced great difficulties gaining institutional acceptance of PSI. Skinner was not unaware of administrative inertia. In his fictional community (described in Walden Two, 1948), he explained how human problems could be solved by behavioural technology. An essential step in the establishment of this community was to create it separate from the existing
community so as to avoid the constraints of the old framework. This was a luxury which programmed instruction did not have.

Skinner suggested that cultural inertia worked against Pressey’s device, but it also appears to have had a role in the downfall of programmed instruction: The difficulties associated with humanists, teachers, and institutions, are all aspects of this inertia (also see Benjamin, 1988). It is important to note, however, that this explanation has been dismissed by at least one writer (Buck, 1990). Buck argues that the inconsistent findings from media comparison studies imply that educators were not convinced of the superiority of teaching machines (and presumably, programmed instruction) over traditional methods and for this reason did not adopt the approach. Buck adds that the only need for either of Pressey’s or Skinner’s machines was one perceived by the creators. However, the large financial, research, production, and marketing commitments to programmed instruction and teaching machines suggests that many financiers, educators, and researchers did see great potential in this approach.

Summary of Skinner’s Contribution

Oprant research provided several principles of learning that Skinner then applied to education. According to this application, learning is promoted by active participation, immediate reinforcement of responses, systematic progression of materials from simple to complex, providing external assistance for responding in early learning but gradually fading this support, arranging materials so that students cannot progress without participating, and promoting complex discriminations by using a variety of examples and discussing concepts from different perspectives. In addition, Skinner advocated that students should control their own study pace and that instructors should not use aversive methods of control. Instrumentation was used to bring learners into contact with these contingencies.

Although the impact of programmed instruction was large and sustained, it eventually failed for a number of reasons. It is not apparent that ineffective technique was among these. On the contrary, programmed instruction appeared to have strong theoretical and empirical credentials that could form the basis of further development of techniques that promote learning (PSI is one such example). The movement left two important legacies for future technical innovations in education: (a) Educational behaviour could be analysed and governed according to the same psychological principles that govern other behaviours; and (b) It is the underlying psychological concepts implemented using technology, rather than the technological vehicle, that is important in influencing student behaviour.
Chapter 1

The following sections will consider the modern version of automated instruction, CBI, and some related instructional issues. Because the computer is an important and ubiquitous tool in instructional settings, measures of its effectiveness in education will be reviewed. It is argued that, as an instructional tool, the computer has failed to meet expectations. One of the reasons for this has been an overemphasis on displaying the technical capabilities of the device, and a lack of utilisation of fundamental psychological principles in the software. That is, it has promoted the technological vehicle, and neglected the principles of learning so strongly emphasised in programmed instruction.

Following a consideration of some problems in CBI, the focus will be turned to the issues of learner control and, in particular, the long-standing tenet of providing the learner with pacing control. One area in which the capabilities of the computer have been exploited is in providing learners with control over a range of instructional decisions (e.g., how quickly to progress, which topics to study). Although this demonstrates the flexibility of the device, there is not yet sufficient evidence that instructional choice increases learning. The most common form of instructional choice is allowing students to determine their own study pace. Self pacing was advocated by both Pressey and Skinner, and it continues to be the dominant approach in CBI, but there has been little empirical investigation of the relative merits of self pacing and external pacing (i.e., determined by instructional designers) on instructional tasks. Evidence from a variety of learning tasks will be presented that demonstrate that learning is not optimal under self pacing, and that it can be improved by external pacing. The experiments of this thesis will then present an empirical investigation of the effects of self pacing and external pacing on learner performance as students complete programmed instructional materials.

Computer-Based Instruction (CBI)

A Typology of Computer-Based Instruction (CBI)

Computer-based instructional software (also known as courseware) comes in many varieties, but is typically labelled with generic terms such as computer-based instruction, computer-assisted instruction, computer-based teaching, computer-based education, or computer-assisted learning.

The actual term used to describe the use of computers as an instructional device is often determined by the author's preference. Many authors use different terminology, and there appears to be little standardisation in the literature. The present thesis employs the term computer-based instruction (CBI) because it is sufficiently broad to encompass the diverse uses of computers in instruction (i.e.,
simulations, drill and practice, tutorials, modelling, and progress management). Any particular piece of courseware may have more than one of these features, however, and, the use of some form of progress management or performance record is common in many applications. Hence, it is difficult to categorise different types of CBI.

Table 1-2
One Classification of Computer-Based Instructional Applications
(C.-L. C. Kulik & Kulik, 1991)

<table>
<thead>
<tr>
<th>CBI (computer-based instruction)</th>
<th>CAI (computer-assisted instruction)</th>
<th>CMI (computer-managed instruction)</th>
<th>CEI (computer-enriched instruction)</th>
</tr>
</thead>
<tbody>
<tr>
<td>tutorial</td>
<td>drill and practice</td>
<td>simulation</td>
<td>programming problem solving</td>
</tr>
</tbody>
</table>

The Kuliks and their colleagues who investigated the efficacy of different forms of computer assistance (e.g., C.-L. C. Kulik & Kulik, 1991) have used CBI as an umbrella term which encompasses three classes of courseware: computer-assisted instruction (CAI), computer-managed instruction (CMI), and computer-enriched instruction (CEI).

In this classification (see Table 1-2), the first branch of CBI is CAI which is typified by drill and practice exercises and tutorial instruction. Drill and practice programs assume some prior knowledge of a subject area, and drill the student on this knowledge. Typically a question is presented and is accompanied by multiple-choice alternative answers. After students respond, they are informed of the correctness of their answer, and may be provided with some additional remedial information. The tutorial is a variation on this theme in which instructional material is presented prior to questions. Programmed instruction as promoted by Skinner, and which will be employed in the studies of this thesis, is an example of tutorial instruction.

CMI refers to the computer's control over the instructional route taken by a learner. The management system may provide a test to assess entering behaviour,
provide a diagnosis of weak and strong areas in a student's repertoire, direct the student to relevant learning resources (e.g., books, CAI lessons), and, most commonly, maintain a comprehensive record of student activity. Teachers can use information provided by the computer to assist their instructional decisions. Most types of CBI include some form of CMI.

CEI encompasses simulations, programming devices, and problem solving programs. Simulation programs present a model (via graphics or animation) which is analogous to some real system. Variables that are effective in the real system can be manipulated in simulations to enable an understanding of the system's operation. In this way students can, in a sense, utilise equipment which is dangerous, delicate, expensive, or not readily available. As a programming tool, the computer compiles and executes programs devised by the student. In a problem solving role, the computer operates as a tool for obtaining the information necessary to complete a problem (e.g., calculating device).

The variety of elaborate presentation modes available in CBI could, ideally, provide learners with the instructional style and content that is needed at any particular point in the learning sequence. CBI's success in improving student performance will be evaluated after a brief consideration of the origins of CBI.

**Origins Of Computer-Based Instruction**

Computers and CBI are the modern versions of teaching machines and are expected to succeed where their predecessors failed. When first introduced, there were obvious benefits associated with digital computers: they were powerful, endlessly patient, had a large, permanent storage capacity, and, according to Skinner, were a smoother operating version of the mechanical teaching machine (cited in Green, 1984). In addition, computers had a versatility previously unseen because programs could be modified very quickly and delivered without the production problems of nondigital materials. As instructors, computers promised to assess students needs, prescribe materials, monitor progress, and, based on student performance, adjust instruction precisely to individual student needs.

Expectations about CBI's influence on student performance have not been conservative; on the contrary, extravagant claims have been made about their contribution. When computers were first introduced as teaching devices, they were considered to offer limitless potential, to be the beginning of a new epoch in education, and likely to have as profound an effect on education as the use of writing tools and the printing press did before them. Computers were expected to be a magic cure for all that ailed education. Expectations of these sort were not entirely new. Educators have always attempted to use innovative technology (e.g.,
printed word, film, radio, television, programmed instruction) profitably, and computers were the next in a long line of educational technologies. Nevertheless, the flexibility of computers gave the impression that they would fulfill the promise that previous technologies had not. The reality, however, has been quite different, and this will be discussed in more detail below.

Computer-based instruction was first used by the computer industry in the late 1950's as a means of training personnel. Early programming efforts required the use of machine language, however, which was difficult to understand. It was not until IBM provided software to assist in the development of instructional programs in the 1960's that education became more involved with computers (Suppes & Macken, 1978). Since the late 1960's many commercial, corporate, and educational institutions have spent large amounts of time and money developing instructional software (see Baker, 1978; Hofstetter, 1985; Kearsley, 1976; Pagliaro, 1983; Suppes & Macken, 1978, for descriptions of some early projects and activity in the CBI field).

In the 1960's and 1970's, the National Science Foundation invested over $14 million in the development of two systems which were largely responsible for the growth of CBI: PLATO (Programmed Logic for Automatic Teaching Operations) and TICCIT (Time-shared, Interactive, Computer-Controlled, Information Television). PLATO, developed during the 1960's, was used to teach many topics including mathematics, chemistry, biology, and accounting. Instruction was delivered using drill and practice, tutorial, and simulation formats. The software used to generate programs in PLATO had no underlying instructional strategy built in (Sugarman, 1978), so instructors could generate programs using instructional features that suited them. In terms of hardware, PLATO consisted of a network of time-sharing terminals linked to a large host computer. To interact with the computer, students used a keyboard with both alphanumeric and special function keys (e.g., help), and a touch sensitive display unit. Other peripheral devices such as slide and film projectors could also be connected at the site and controlled by the computer. Some advantages of this system were that the host could maintain a large library of programs for student use, use powerful programs, and service many students at one time; for example, one central machine based at a university could service hundreds of remote sites and thousands of terminals at those sites. In accordance with changes in technology, PLATO systems changed over time (Niemiec & Walberg, 1989; Pagliaro, 1983) and have continued to supply CBI to many students. In more recent years, microcomputer technology has
been combined with the PLATO system (micro-PLATO) in an attempt to derive the benefits of both hardware and software development (Eddins, 1982).

TICCIT (Merrill, 1980) was developed in the 1970's, and considered an advance on previous forms of CBI because it was an application of an instructional theory (Reigeluth, 1979). TICCIT was reported as "the first time, since the application of Pressey's simple mastery paradigm and Skinner's immediate reinforcement to the early teaching machines, that CAI formally and rigorously attempted to base its operations in instructional theory" (Pagliaro, 1983, p. 80). Instruction in TICCIT was built around the notion that ideas should be presented in three modes; rules, examples, and practice. Materials were presented in these modes primarily with diagrams and words. Using control keys, students were able to select such things as pace, topic, difficulty, and mode of presentation. This control was designed to promote independent learning. To deliver instruction, TICCIT used a host-terminal system and utilised closed-circuit television as a display device. Students using TICCIT saw a television screen and a keyboard that had both the traditional keys plus additional keys for editing responses and controlling various aspects of the instruction (e.g., practice, examples, advice).

Field evaluations of PLATO and TICCIT (Alderman, 1978; Alderman, Appel, & Murphy, 1978; Murphy & Appel, 1977) did not produce flattering results for either system: Compared with traditionally-instructed classes, PLATO did not increase student achievement, and TICCIT did increase achievement but also increased withdrawal rates.

Time-sharing computer systems had many restrictions that limited their feasibility as teaching devices. They were expensive. Because they were based around a large, single central processor users could experience long response delays at times of heavy usage, access could not be guaranteed, and, when the system broke down, all users were affected. A disadvantage of computers advanced less frequently was the "psychological threat" they posed for the student (Dick, 1965); this threat has been relabelled such things as "computer anxiety" and "compuphobia" in more recent years (Torkzadeh & Angulo, 1992).

The advent of the microcomputer changed things radically. Due to advances in microcircuitry and silicon chip technology, microcomputers were developed which were as powerful and flexible as their mainframe predecessors. In the late 1970's fully-assembled microcomputers became available and, in 1981, the IBM PC was released (Hofstetter, 1985). This machine became an icon of the microcomputer industry. Now, users could purchase an off-the-shelf computer with the power of a mainframe, but which was many times smaller and cheaper.
They were quick, standalone, provided guaranteed access, and released users from "the tyranny of the shared machine" (Howe, 1978, p. 121). Hence, most of the problems associated with mainframes had been solved by microcomputers. The problem of computer anxiety, however, continues to exist in many end-users and its definition, aetiology, and effects are subjects of ongoing research (Torkzadeh & Angulo, 1992).

The advantages of microcomputers, particularly their low cost, made CBI more attractive to educators (Eisele, 1979); there was an expectation that CBI would not only increase instructional outcomes but that it could do so at a reduced cost. The educational market is massive, and CBI developers were certainly aware of this. Conditions were ripe for a large expansion in educational computing. In the late 1980's it was estimated that 90% of all American schools used computers in instruction (Niemiec & Walberg, 1987). By 1990, some regarded educational computing to be almost as common as the chalkboard (Chaparro & Halcomb, 1990). Survey figures regarding computer use in college psychology instruction supported these claims (Anderson & Hornby, 1990). A survey was mailed to 135 institutions, primarily in the U.S., but some were in Canada and outside North America. Responses were obtained from 72 institutions (53%) and did not differ according to geography. Over 80% of respondents worked in a department which used computers in psychology instruction. Of the respondents, 66% indicated that their Psychology Department had their own microcomputer laboratory (with a mode of 12 machines), and 55% of the departments were aiming to increase the number of computers. The authors also predicted future increases in the number of psychology department microlaboratories. Following sections consider the use and effectiveness of this large number of computers dedicated to education.

Effectiveness of CBI

Studies assessing the relative influence of CBI and traditional instruction on student performance have been so abundant that some reviewers of this area have proposed that additional comparative experiments would not improve the existing understanding of the effectiveness of CBI (Niemiec & Walberg, 1985). Because of this abundance of information, and, due to the development of quantitative analytic techniques during the time of computers, meta-analyses of findings from comparison studies have become popular. Furthermore, their findings have become a major source of support for proponents of CBI. It is not the task of this thesis to consider each of these reviews closely, but rather to present an overview of their findings and consider how these findings can be interpreted.
The outcome variable that has received most attention is examination score following completion of instruction. Did the group that used CBI perform better, on average, than the group receiving traditional instruction? Table 1-3 summarises the findings of several meta-analyses published in the last decade. Most reviews presented in the table come from a central group of researchers (the Kulik's and their colleagues) whose technique has been to collect individual media-comparison studies (i.e., experimental comparisons of one form of instruction with another) and, using one such study as the unit of analysis, calculate an average ES due to CBI. This is the typical meta-analytic approach. The review of Niemiec and Walberg (1987) is unusual because they used the meta-analytic review (as opposed to one study) as the unit of analysis. Hence, their data summarises numerous metanalytic findings.

There are two important kinds of data provided by the reviews in Table 1-3: ES’s and frequency counts (both described previously). Because they allow different interpretations of the typical, or representative, effect of CBI, they will be described separately.

**Effect Sizes**

The column of mean ES’s in Table 1-3 shows that each review found a positive ES (on average, exam achievement scores were higher in CBI than traditional instruction), and that the range of improvement was from 0.26 to 0.47 standard deviations. To place the effect sizes in some perspective, a value of 0.26 means that the average effect of computers was to raise examination scores by 0.26 standard deviations, or from the 50th to the 60th percentile. A value of 0.47 represents a shift in performance from the 50th to the 68th percentile. Niemiec and Walberg’s synthesis indicates that CBI can typically be expected to shift performance from the 50th to the 66th percentile. These gains are similar to, or greater than, those found in reviews of programmed instruction. They represent a significant improvement in performance, but not one that is overwhelming; especially in light of the expectations that have accompanied CBI.
Table 1-3
Percentage of Studies in Which the Effect of Computer-Based Instruction (CBI) and Traditional Instruction (TI) on Achievement was Statistically Equivalent, or Statistically Significant in Favour of One Method, Plus the Mean Effect Size

<table>
<thead>
<tr>
<th>Reviewers</th>
<th>Number of studies&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Level of student&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Mean effect size</th>
<th>CBI = TI (%)</th>
<th>CBI &gt; TI (%)</th>
<th>CBI &lt; TI (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-L. C. Kulik &amp; Kulik (1991)</td>
<td>248</td>
<td>K to adult</td>
<td>0.30</td>
<td>60</td>
<td>38</td>
<td>2</td>
</tr>
<tr>
<td>Niemiec &amp; Walberg (1987)</td>
<td>13 reviews</td>
<td>mixed</td>
<td>0.42</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Niemiec, Samson, Weinstein, &amp; Walberg (1987)</td>
<td>48</td>
<td>elementary</td>
<td>0.45</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>J. A. Kulik &amp; Kulik (1987)&lt;sup&gt;c&lt;/sup&gt;</td>
<td>199</td>
<td>mixed</td>
<td>0.31</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>C-L. C. Kulik, Kulik, &amp; Shwalb (1986)</td>
<td>23</td>
<td>adult</td>
<td>0.42</td>
<td>52</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>C-L. C. Kulik &amp; Kulik (1986)</td>
<td>99</td>
<td>college</td>
<td>0.26</td>
<td>78</td>
<td>21</td>
<td>1</td>
</tr>
<tr>
<td>Bangert-Drowns, Kulik, &amp; Kulik (1985)</td>
<td>42</td>
<td>secondary</td>
<td>0.26</td>
<td>57</td>
<td>38</td>
<td>5</td>
</tr>
<tr>
<td>J. A. Kulik, Kulik, &amp; Bangert-Drowns (1985)</td>
<td>28</td>
<td>elementary</td>
<td>0.47</td>
<td>18</td>
<td>82</td>
<td>0</td>
</tr>
</tbody>
</table>

Note. A dash (-) indicates that this information was not available from the article.

<sup>a</sup> This figure represents the number of studies analyzing final exam achievement. Because this is the primary outcome measure in most analyses, this figure closely approximates the total number of studies reviewed. <sup>b</sup>K = Kindergarten. Mixed = one or more of elementary or secondary level, college, exceptional populations, nontraditional adult education (e.g., flight training, electronics). <sup>c</sup>This paper is a summary of the authors' four previous meta-analyses.
Frequency Count

Although ES’s have the advantage of being a single, standard score that represents the average effect of a treatment, they can have the distinct disadvantage of being a misleading indicator of the average or typical outcome from comparison studies. Columns 5, 6, and 7 in Table 1-3 present the frequency counts from the same reviews. Column 5 (labelled “CBI = T1”) shows that the most common finding in the studies used for these meta-analytic reviews, was no significant difference between the computer-instructed and traditionally-instructed groups. Typically, fewer than half the experiments found any significant differences (Columns 6 and 7); when they did, however, CBI generally led to better performance.

Meta-analyses often report that CBI raises student achievement in the average study, but this is misleading because the ES often does not represent the typical outcome of an original comparison study. These averages refer to a mean finding in which CBI is benefited by many studies that find small but insignificant advantages for this method, or a smaller number of studies that find atypically large advantages for this approach. The modal outcome in these comparative studies is that there is no significant difference between the achievement resulting from CBI and traditional instruction. This information is lost in the ES.

Media-comparison studies (e.g., CBI vs. traditional instruction) have been criticised for their over reliance on achievement outcomes, and lack of emphasis on the evaluation of education beyond the final exam score (J. A. Kulik’s views are presented in Bonham, 1990). For example, some other issues of importance are whether student attitudes or feelings are also altered by alternative teaching approaches, and whether the technique is time and cost effective. Hence, it is relevant that meta-analyses have found that CBI is often associated with positive changes in students’ attitudes towards computers and towards quality of instruction (J. A. Kulik & Kulik, 1987; C-L. C. Kulik & Kulik, 1991; Niemiec, Samson, Weinstein, & Walberg, 1987). In addition, CBI students typically require about one-third less instructional time than students learning from traditional instruction (J. A. Kulik & Kulik, 1987; C-L. C. Kulik & Kulik, 1991), and CBI is a cost-effective means of improving performance (Niemiec, Sikorski, & Walberg, 1989).

In summary, meta-analytic reviews indicate that CBI and traditional instruction are often not different in their effectiveness on achievement. When CBI and traditional instruction do differ, however, the difference favours CBI. In addition, CBI can have affective and time benefits.
Do Computers Contribute Anything to Learning?

Findings from meta-analytic reviews can be challenged on the grounds that average outcomes are often not representative of most studies. There are also other serious problems with interpreting findings from the comparative studies used in these reviews. Before concluding this section on the influence of computers on learning, it is important to consider a view which has existed for many years (e.g., Schramm, 1977), and which has been reiterated recently with respect to educational computing (Clark, 1982b, 1983, 1985a, 1985b, 1991, 1992).

In a series of provocative articles Clark has maintained that media (e.g., television, radio, teaching machines, computers) are not responsible for learning under any circumstances, and argues that “media are mere vehicles that deliver instruction but do not influence student achievement any more than the truck that delivers our groceries causes changes in our nutrition” (1983, p. 445). His point is that the function of computers in an instructional medium should not be confused with the active instructional ingredients that may covary with it. In particular, Clark considers instructional method, content, and novelty to be uncontrolled confounds that account for the advantages generally attributed to computers.

Instructional Method

In a reanalysis of a sample of studies used by J. A. Kulik and his colleagues, Clark (1985b) found that when the same teacher presented both CBI and traditional instruction, the effect size of CBI reduced from significant to nonsignificant levels. Clark’s findings support an account in which instructional method and content are the active ingredients of CBI. That is, when only one teacher is used to present both CBI and traditional instruction, those methods are more likely to be equivalent on critical instructional elements (e.g., clarity of objectives, careful preparation and organisation of the materials, content, examples, practice, pacing, feedback), and the no-difference-between-treatments outcome is accurate. Furthermore, when method and content were statistically controlled in the reanalysis, significantly fewer studies favoured CBI than when these controls were not in place. This pattern of results continues to occur. Recent meta-analyses (C-L. C. Kulik & Kulik, 1991; J. A. Kulik & Kulik, 1987) have shown that there is a statistically significant decline in effect size (from about 0.4 to about 0.25) when instructor effects are controlled. Hence, the results of meta-analyses support the notion that instructional method is an influential confound that accounts for at least some of the effects commonly attributed to computer use.
Novelty

Novelty provides a second rival explanation of the instructional benefits generally attributed to computers (Clark, 1985a). There is evidence that the longer a study progresses, the smaller the achievement differences between the computer-instructed, and traditionally-instructed, groups become. In an early meta-analysis (J. A. Kulik, Bangert, & Williams, 1983) it was found that the ES of CBI decreased systematically as study duration increased: from 0.56 to 0.3 to 0.2, for durations of four weeks or less, 5 to 8 weeks, and more than 8 weeks, respectively. In a more recent meta-analysis (C-L. C. Kulik & Kulik, 1991), studies lasting 1-4 weeks had a significantly higher ES (0.42) than studies lasting 5 weeks or more (ES = 0.26). This effect occurred in both precollege and postsecondary samples.

Editorial Gatekeeping

The face validity of CBI also may have benefited from publication bias. Achievement effect sizes are consistently found to be two or three times larger in published studies than in unpublished ones, and the average effect size for the latter studies is quite modest (C-L. C. Kulik & Kulik, 1991; J. A. Kulik & Kulik, 1987; Niemi & et al., 1987). If researchers, editors, and publishers do have a publication bias towards large positive results, then reviews of comparative studies, and the literature in general, give an exaggerated impression of CBI’s true worth. Because this impression typically includes incorrect assumptions about the active ingredients in the media effect, it is doubly misleading.

Some reviewers consider the differences between published and unpublished studies to be the tip of the iceberg, that many more findings never see the light of day, and that results from published studies should be discounted (J. A. Kulik & Kulik, 1987). The authors add, however, that differences in resources, experience, and circumstances may exist between researchers who are published and those who are not. These differences will probably affect the quality and findings of the study. Nevertheless, the methodological confounds and editorial gatekeeping concerns are serious enough to postpone any definite conclusions regarding CBI’s superiority over traditional instruction.

Other Factors

Evaluating the impact of CBI is further hindered because study outcomes are inconsistent. Replications, or near replications, produce different results, and attempts to correlate study features with outcomes have only been partially successful at identifying influential variables. Determining why different studies of
the same issue often lead to different results remains a key problem for researchers in this area (J. A. Kulik, cited in Bonham, 1990).

A factor that appears to modify results from CBI comparison studies is the educational level of the student. It has been argued (J. A. Kulik, 1981) that achievement under CBI is an inverse function of educational level, and there is some evidence to support this: CBI is most effective at the elementary level, least effective at college level, and is of intermediate effectiveness, but at the lower end, for the secondary level (see Table 1-3, and Niemiec & Walberg, 1987). This may be overly simplistic, however, because it suggests a universal effectiveness of CBI at each educational level. In fact, there is also considerable evidence for an interaction between type of CBI and educational level; the effectiveness and appropriateness of CBI use at each instructional level seems to be important. At the elementary level, drill and practice has been found to be effective, but CMI has not. At high school, both CAI and CMI have had good effects, but CEI has not. At college level, each type of CBI has had similar effects. These findings may be explained if younger students benefit from the good structure and small steps offered by CAI, and older students are less reliant on structured materials and profit more from the independence and choice of CMI (Bangert-Drowns, Kulik, & Kulik, 1985). A similar account may explain why low achieving students gain more from CBI than higher achievers (J. A. Kulik, Kulik, & Bangert-Drowns, 1985; Niemiec et al., 1987).

The view that computers do not contribute to learning is both counterintuitive and unpopular, but it is important to note that it has recently been reported (Clark, 1991) that J. A. Kulik agrees that it is the teaching method built into CBI that accounts for its positive influence. This supports the notion that research efforts could be more profitably invested in instructional (cf. media) issues.

Conclusions from Meta-Analytic Reviews

Review findings. Meta-analytic reviews of studies comparing CBI and traditional instruction have consistently indicated that the use of computers is advantageous for learning. The data in Table 1-3, however, indicate that the ES, which is employed to represent a typical outcome, does not necessarily reflect the most common finding. Most often, comparative studies find no significant differences between CBI and traditional instruction. Furthermore, Clark’s reinterpretation of the meta-analytic data dealing with CBI is logical and cogent, and indicates that comparison studies have problems with internal validity and inconsistent findings (described above) which also make conclusions unreliable. Nevertheless, media enthusiasm for CBI has not waned, probably because counter
evidence takes a long time to be disseminated (Clark, 1983), and a multi-million dollar computer industry (including software and hardware producers, and sales personnel) has a vested interest in maintaining computer enthusiasm (Clark, 1991).

**Media-comparison studies.** Some general points need to be made about media-comparison studies in general, and how they relate to the research that is presented in the experiments of this thesis. With the advent of each new instructional medium (e.g., programmed instruction, TV, instructional radio, PSI, CBI), comparison studies are made in a process that resembles reinventing the wheel. A class using the new method is compared on some performance measure to a class instructed in a traditional manner, and an attempt is made to use the findings to argue which is the better instructional medium. As was demonstrated using CBI in the previous section, media-comparison studies fail to acknowledge fundamental differences in things such as pacing, control, and interactivity, for example, between the approaches being compared (Foster & Booth, 1993; Lumsdaine, 1962; Morrill, 1961; Schramm, 1977). There are two implications of this: (a) the findings are of limited usefulness in the evaluation or development of instructional techniques, and (b) because media-comparison studies cannot generate clear evidence that a new approach is superior to traditional methods, the potentially effective instructional approach may lose research, educational, and public enthusiasm.

Perhaps the exercise of comparing instructional media is misguided. Other approaches to evaluating the merits of teaching approaches have defined accomplishment in behavioural terms (e.g., what size vocabulary can be learnt, what kind of problems can be solved, at what rate, and recalled for how long?). Once the objectives are defined, efforts can be put into producing and trialing methods that meet these goals most effectively (e.g., Schramm, 1977; Taber, Glaser, & Schaefer, 1965).

**CBI.** CBI research may be more productive if it is directed towards pedagogical issues rather than comparative studies. CBI would benefit from an increased investigation of the specific software features that contribute to learning (Karrer, 1991), and by an increased experimental rigour and specificity in the primary research (Niemiec & Walberg, 1987). These were important considerations in the experiments of this thesis.

**Quality Problems in CBI**

Whereas the previous section considered comparative studies and statistical attempts to determine whether, and why, CBI is educationally effective, the present section presents the views of psychologists, educators, and scientists who have
been disappointed with CBI and its inability to meet expectations. In particular, the
distinction between the relative contribution of digital and pedagogical techniques to
CBI, that was highlighted by Clark, is extended.

Educational Software: Technology and Psychology

Many educators are disillusioned with educational software. CBI's inability
to meet its potential as an educational remedy has been attributed to much of the
software being of poor quality (Barrett, Micceri, & Pritchard, 1990; Winship,
1992), reflecting poor pedagogy (Fisher, 1982), containing serious instructional
flaws (Vargus, 1986; Vargus & Vargus, 1991), being badly written (Hewett,
1988), poorly developed and inadequately documented (Hornby & Anderson,
1988), and lacking the fundamentals of effective instructional design (Merrill,
1988).

Emphasis on Technology. It is paradoxical that in CBI the relatively new
digital medium (i.e., the hardware) is much more sophisticated than the relatively
old techniques of psychology and education. This paradox has been summarised
neatly as follows: “Microcomputers are an integral part of our educational system.
This phenomenon has occurred despite the fact that the technological sophistication
of the machine far exceeds the technological sophistication of the teaching
procedures most often used in educational software” (LeBlanc, Hoko,
Aangenbrug, & Etzel, 1985, p. 23). The CBI industry’s emphasis on technology
has been described as “technocentric” because developers have attempted to
“optimize the capabilities of technology rather than learners” (Hooper & Hannafin,
1991, p. 70). This is obviously a flawed approach in education where the measure
of success is the quality and quantity of change in the student.

Problems with Software Quality. The Educational Products Information
Exchange (EPIE) is an independent, non-profit agency that evaluates teaching
materials and equipment. After a review of hundreds of programs, they concluded
that “only 5% of these hundreds of programs have been judged to be of truly high
quality, while more than half have been judged not worth recommending to
educators or parents” (Komoski, 1984, p. 247). In a subsequent report indicating
the bases of software problems they state

There was an overwhelming lack of field testing evidence in the course of
program development. Many programs had no support materials, unclear or
developmentally inappropriate learner objectives, and few instructional
suggestions or information to aid in curricular integration. Most programs
failed to use an approach which lent itself to an effective delivery. There was
little evaluation of student learning and few programs included a management system. (Bialo & Erickson, 1985, p. 227)

These are obviously crucial omissions from the development of software and many appear to be eminently fixable. There is some evidence of improvement in software quality: In ongoing courseware reviews by the EPIE (Sivin & Bialo, 1988), the ratio of recommended to reviewed programs increased post-1985 compared with pre-1985, leading the authors to conclude that quality "seems to have improved" (p. 57). There is also evidence, however, that quality has not improved. Dudley-Marling and Owston (1987) reported that only 5% of 105 software evaluations by a panel of teachers were rated exemplary in terms of pedagogical content and instructional presentation, whereas nearly 30% were rated deficient on each of these measures. These figures had barely changed over a number of years (see Owston & Dudley-Marling, 1988). Similarly, a meta-analytic review of comparative studies of CBI and traditional instruction, conducted between 1966 and 1986, found that ES's did not increase over that period (C-L. C. Kulik & Kulik, 1991).

More recently, Barrett et al. (1990) likened technology-mediated instruction to a dog and pony show in which instructional design and content take backseat to flashy presentation. In an assessment of over 200 CBI packages they found that few attained even 50% of the optimal score on a measure incorporating physical, presentation, instructional, and management considerations (see also Pritchard, Miccoci, & Barrett, 1989). Some of the criticisms of the software included that "tried and true" instructional techniques (such as drill and practice and problem solving) were rarely used to an extensive degree, that there was very little innovative instruction, that nearly 40% of programs did not employ questions, and those which did typically used a multiple-choice format (greater than 70%), a response format that has been strongly criticised (Bork & Pomicter, 1990; Skinner, 1958). They proposed that much of the publicity accompanying CBI concerns the presence of technology rather than the way it is used. A recent editorial in Psychology Software News expressed the ongoing concern that what psychology has discovered about instruction is rarely integrated into the design of learning systems (Hammond & Trapp, 1993).

Sources of the Software Quality Problem. Producers and consumers both contribute to the software quality problem. Textbook publishers and computer manufacturers often produce educational software. Typically, however, software production is secondary to their main interest, selling books. Educational software
may accompany (and help sell) textbooks, but whether it works may be a secondary concern (Castellan, 1988). Furthermore, software designers who are computer programmers may have expertise in exploiting the medium and using its technological features to interest children but, generally, they do not know much about psychology (Hewett, 1988) or how to teach anything important (Skinner, cited in Zientara, 1984).

Software publishers are becoming more abundant. These companies do not have an alternative market, and their existence depends on product quality, so they are more likely to use extensive review and testing procedures during development (Castellan, 1988). This is a promising shift for both consumers and the field of CBI. An unfortunate obstacle, however, is that economic forces discourage innovative software developers because even if they create successful products, rival publishers can copy the product and even development costs may not be recouped (Sivin & Bialo, 1988).

Much educational software is also written “in house” by academics and researchers who often have an effective blend of instructional and programming knowledge. Even if such products are of better quality than commercial ones, potential consumers may have restricted, or no, access to them because they may not be made widely available, or do not meet contemporary criteria for “good” courseworkware (Eamon, 1988). Not unexpectedly, in-house products can also be trash (Castellan, 1988), and the consumer is left with the ultimate decision of what to use.

Consumers (e.g., schools, parents, businesses) also contribute to the perpetuation of poor courseworkware. Consumers are often naive in their understanding of computers and effective elements of instruction, and if they cannot detect poor software, they are likely to be influenced more by availability and marketing than effectiveness (see Barrett et al., 1990). Hence, they do not pressure software producers to improve their product. Education’s uncritical acceptance of computer technology and the factors that have contributed to this are discussed by Balajthy (1988). Although there have been repeated calls for consumers to pressure producers to increase quality (Barrett et al., 1990; Castellan, 1988; Vargas, 1986), there appears to have been an unfortunate Catch 22 in which neither producer nor consumer is likely to assume responsibility for increasing the standard of courseworkware. When publishers have a product that sells (independent of its effectiveness) they do not need to alter their production practices; consumers, typically, do not have the time, money, or expertise to evaluate the utility of programs, and pressure developers into increasing standards.
It is also true that some dissatisfaction with CBI is attributable to unsuccessful integration of the software into the existing curriculum. Typically, instructional solutions cannot be purchased off the shelf. Even if good quality software exists, it must first be discovered and then integrated, and this is difficult and time consuming (Hammond & Trapp, 1991; Hornby & Anderson, 1988).

**Solutions to the Software Quality Problem.** Various solutions to the software dilemma have been proposed. These include incorporating essential learning principles that are the cornerstone of instructional effectiveness (e.g., Tudor & Bostow, 1991; Vargas & Vargas, 1991), field testing products (see Bialo & Erickson, 1985; Owston & Wideman, 1987), avoiding an over reliance on technological features of computers that are ancillary to the learning process (e.g., Eamon, 1988), devising and utilising objective, reliable, and relevant CBI evaluation tools (see Barrett, et al., 1990; Heller, 1991; Owston & Wideman, 1987; Schueckler & Shuell, 1989), and introducing quality standards such as including development and implementation details on the product (e.g., Castellan, 1988). The experiences of Hornby and Anderson (1988) provide a good example of the frustration, and wasted time and effort that arises from having to trial undocumented courseware first hand to discover what features it utilises (e.g., basic pedagogical devices such as objectives, quizzes). In addition, consumers could raise quality standards by becoming more discriminating buyers, being aware of what elements an effective product should have, and only purchasing after trialing a product and satisfying themselves that it meets their curricular and instructional needs. Unfortunately, although classroom educators are an important group of consumers, in a good position to evaluate the instructional potential of software, they are unlikely to do so because this is a time-consuming activity which provides little professional return (Ellis, 1992; Heller, 1991).

There is no single, or general, solution to the problem of poor quality CBI because there is a variety of software types (see Table 1-2) that differ in aims and approaches. For example, a programmed tutorial is suited to taking a new learner from a level of naivety to relative sophistication in a given topic, whereas an information-heavy database is suited to a discovery approach to learning that is likely to be efficient for advanced learners who know what they are looking for. Nevertheless, for tutorial materials that are designed to teach new information to relatively naive learners, there are established principles and techniques for effective delivery that have been researched and articulated for many years (e.g., Geller, 1992; Vargas, 1986; Vargas & Vargas, 1991). These techniques, however, have been paid only lip service in traditional instruction, and to repeat this in CBI is to
lose an opportunity to make it a truly effective approach (Foster & Booth, 1993). There has been little research on how psychological principles of learning operate in CBI (Shuell & Schueckler, 1989), perhaps because the preoccupation has so far been with “computer-based” rather than “instruction”.

CBI: Summary And Future Research

Computers appeared to offer the solution to many problems of effective individualised instruction. To evaluate the effectiveness of CBI many studies compared student achievement after being taught by computer or traditional methods. Meta-analytic reviews of these studies showed that, overall, CBI raised student achievement, but the amount of improvement was not large, and many of the original studies found no significant differences due to instructional method. Furthermore, those media-comparison studies appear to have been confounded because alternative explanations of CBI effectiveness, such as instructional method and novelty, seem more reasonable. Because these studies cannot control key instructional factors (such as rate of learner progress, amount of interaction with materials) they have been criticised for not being sufficiently specific and rigorous. Therefore, they do not assist understanding of what factors contribute to effective implementation. CBI has also been criticised for emphasising the computer’s bells and whistles while failing to utilise psychological principles to facilitate student learning.

CBI would benefit from a reduced emphasis on media-focused studies, and increased research attention to pedagogy. For instance, understanding when drill and practice is more effective than other forms of CBI, or whether incorporating response time limits in programs improves performance, are the types of psychological issues important to both psychologists and educators.

One important instructional issue borne out of CBI, is the extent to which students control various aspects of their own instruction (e.g., rate of progress, sequence of materials, amount of practice). Although computers can be programmed to give students a wide range of control (from none to almost unlimited), there are few experimentally-based prescriptions for how much control should be provided. The following sections describe the pros and cons of learner control in general before focusing on one aspect of control, rate of progress through materials. A disparity between long-held prescriptions for who should progress and relevant experimental evidence is then considered; this is the basis for the experiments of this thesis.
Learner Control

Teaching machines were mechanical devices designed to present linear and preprogrammed sequences of instruction. Although determination of which materials to present, and in what fashion, was made by the subject matter expert and programmer, this type of instruction could be considered individualised because each student received plentiful reinforcement contingent on their behaviour, and they worked at their own pace. A predetermined instructional pathway has often been criticised, however, as being restrictive and not sufficiently individualised. One way in which the computer has been expected to increase learning is by enabling an increased number of instructional elements (subject matter, amount of practice) to be placed under the control of the student.

In education it is axiomatic that learners should receive instruction that is tailored to their needs. In other words, instruction should be individualised. A desirable part of such instruction is believed to be learner control (Chung & Reigeluth, 1992). There are at least two reasons for this. First, providing learners with control over their instructional environment should be motivating. The notion has intuitive appeal: Learner control should alleviate boredom, frustration, and anxiety by enabling students to avoid material that they have already studied, or do not want to study. Hence, this logic implies that “learner control will maintain attention longer, involve students more deeply, and perhaps give students greater insights” (Steinberg, 1989, p. 117). Personal control should make an individual feel competent and self determining, resulting in intrinsic interest in the activity and, ultimately, better performance (Kinzie, Sullivan, & Berdel, 1988). The second reason concerns CBI’s apparent advantage of flexibility. Used in one way, the system would assess new students’ current abilities and deficiencies, and then tailor information and presentation to teach them optimally. One reaction to this type of “spoon feeding” is that it could impair learning in a real world that is not so accommodating. In this world, people must learn from less than optimal information; hence, if we wish to develop students with independent learning skills, they will need to be able to adapt the learning system to themselves rather than have it adapt to them (Merrill, 1988). Reigeluth and Stein (1983) also suggest that effectiveness and efficiency of instruction should increase as the degree of learner control increases. Hence, education should produce students who can manipulate their own learning environment.

There has been continued controversy over which instructional elements should be placed under the control of students (internal control), and which should be determined by instructional designers (external control). These elements include
Chapter 1

selecting the content of materials to be studied, sequencing and pacing of these materials, and amount of elaboration and reviewing. Although providing the student with complete control over the instructional presentation is one of the basic premises of current CBI (Kline, 1992), the proposed motivational and performance benefits of learner control have not been well substantiated empirically (Higginbotham-Wheat, 1990; Steinberg, 1977, 1989): In fact, findings have more often been negative than positive (Chung & Reigeluth, 1992).

One problem with giving students control over their instruction is that they often cannot adequately monitor the quality or quantity of instruction necessary for good performance (Carrier, 1984; Chung & Reigeluth, 1992; Clark, 1982a). For example, control of pacing can lead to procrastination, control of sequence can fail because it requires prior knowledge of the conceptual order of materials, control over the amount of practice and level of difficulty can fail if students terminate prematurely because they cannot assess their own level of comprehension. “It is a rare student who knows exactly what he or she should be spending time studying. Most want and need at least some structure and guidance” (Ejser, 1988, p. 46). Without this assistance, self-controlled learners may retain less information than students under external control (Chaparro & Halcomb, 1990).

The negative outcomes that can occur when giving control to naive learners have been demonstrated by Gay (1986). In this study, students were classified as low or high in conceptual understanding of protein synthesis based on their pretest scores. Each of the two groups was then further divided into a learner-control (learners determined their own sequence, amount of practice, presentation mode and content) or program-control condition (program controlled all instructional elements). Pacing was always learner controlled. Posttest results showed that subjects with low conceptual understanding who controlled their study performed significantly worse than the other three groups which did not differ (although a possible ceiling effect may have restricted differences). So, when understanding was good, it made no difference to performance whether control came from the learner or program. When understanding was poor, however, program control was beneficial, but learner control was not. Gay’s study showed that learner control can be used without loss of performance by some students. In fact, the high-conceptual-understanding group who used learner control required significantly less time to complete the materials. More sophisticated learners can sometimes use control options effectively, but there is also evidence that they will not necessarily do so (Clark, 1982a).
There have been successful attempts to individualise control by providing a compromise between complete external control and unrestricted learner control. For example, advisement strategies involve using the program to diagnose a learner's needs and to prescribe a sequence and quantity of study for the student. The student, however, makes the decision regarding whether to follow the program's advice (see Steinberg, 1989, for a brief review of advisement strategies and other tools to assist learners). The advantage of this approach is that learner performance may be increased without removing the individual's sense of control over their study. It has been suggested, however, that advisement assists high-ability students more than low-ability students, and that this shows merely that superior performance is a result of existing study strategies rather than the control provided (Higginbotham-Wheat, 1990).

A different form of compromise between internal and external control involves maintaining program control of instructional support, while giving learners control over contextual variables (such as text density, and problem context) which do not decrease instructional support. This approach offers both the instructional benefits of program control, and affective benefits of learner control (Higginbotham-Wheat, 1990; Morrison, Ross, O'Dell, & Schultz, 1988).

Although there are continuing calls for CBI to provide learners with control over their learning environment (e.g., De Laurentiis, 1993), little reliable evidence exists concerning what elements they should control to improve performance, or the extent of control they should be given. So far, the field of learner control has been more an exploitation of the computer's capabilities than a technique to increase learning.

**Learner Control of Pacing**

One aspect of learner control that has been largely free from controversy is pacing. It has long been an educational ideal that instructional pace should be completely controlled by a learner (self pacing). Self pacing is a distinguishing feature of alternative teaching technologies such as programmed instruction (Skinner, 1954, 1968), PSI (Keller, 1968), the Audio-Tutorial system (Postlethwait, Novak, & Murray, 1964), and CBI. Self pacing is regularly advocated as an appropriate instructional design strategy (e.g., Leshin, Pollock, & Reigeluth, 1992), and it is the component most consistently employed in attempts to individualise instruction (Carrier, 1984).

In their discussion of the software evaluation criteria for assessing tutorials (employed in the EDUCOM/NCRPTAL Higher Education Software Awards), Bangert-Drowns and Kozma (1989) maintain that tutorial software should allow
students some control over instructional presentations (e.g., sequence, pacing, reviews), and should "at least allow students to determine their own instructional pace" (p. 248). They note, however, that students do not always make good self-instructional decisions, so good software will strike a balance between learner control and external control. It seems unclear then, whether these authors advocate complete self pacing. It would seem unwise to take such a position because there is no clear consensus about how it should be employed (Higginbotham-Wheat, 1990). If there should be a balance between learner control and external control of pacing, what is the precise nature of this balance? To determine whether, or to what extent, learners can benefit from external control of pacing, and how this form of control can be implemented, requires empirical investigation. But, "little research has been done to date on the issue of self-paced instruction and micro-computers, probably because of the widely held erroneous assumption that self pacing is the best way to design micro-computer based instructional programs" (Canelos, Baker, Taylor, Belland, & Dwyer, 1985, p. 4).

The self-pacing tenet will be addressed in this thesis by investigating some ways in which a combination of self pacing and external pacing can improve student performance beyond that achieved by self pacing alone. The remainder of this introduction will demonstrate that external pacing of instruction by imposing time delays can be beneficial to learning in a variety of tasks, and that the extent of benefit is determined by the locus of delays and the information available to students during this interval. A strategy for effective use of external pacing will be generated from these findings and tested with programmed instruction; an approach that is based on sound instructional principles, and that is largely responsible for the student self-pacing tenet.

Pacing can be employed at two broad levels: to influence progression between units, or to influence progression within a unit. Much work in the PSI field has shown that self pacing often leads to procrastination that can, in turn, lead to poorer test performance and higher withdrawal rates (J. A. Kulik, Kulik, & Carmichael, 1974; Reiser, 1984). Apart from possible effects on students, procrastination creates difficulties with teaching staff and administrators because students overload the system towards the end of the course or extend their enrolments beyond expected completion dates (Born & Moore, 1978). Hence, it is common for PSI courses to employ some degree of external pacing (such as test deadlines) to avoid problems with student procrastination (e.g., Bijou, Morris, & Parsons, 1976; Born & Moore, 1978; Reiser, 1984). External pacing can be beneficial when employed at the level of unit completion because it promotes
progression from one module of coursework to the next (e.g., Crosbie & Kelly, 1993). The remainder of this thesis will consider how student behaviour is influenced by external pacing applied within each module.

Experimentally-Imposed Delays

Investigation of the effects of experimentally-imposed delays within trials in human learning has a long history. Initially, delaying feedback in simple tasks was a popular procedure because it was an experimental analogue of delaying reinforcement to laboratory animals. Subsequent investigations, however, have considered the influence of interpolating delays between each of the elements in a learning sequence in a variety of complex learning tasks.

Figure 1-5 shows the events occurring in a typical learning task: $S_1$, $R_1$, and $FB_1$ represent the stimulus presented to the subject, their response, and feedback received by the subject (e.g., knowledge of results, praise), respectively; $S_2$ is the subsequent stimulus that begins the next trial. The dashed line between each of these events represents the passage of time and indicates where a time delay can occur. From this it can be seen that a delay can be imposed between a stimulus and a response, between the response and feedback, or between feedback and the next stimulus.

$$
\begin{array}{ccc}
\text{preresponse interval} & \text{prefeedback interval} & \text{postfeedback interval} \\
S_1 \rightarrow & R_1 \rightarrow & FB_1 \rightarrow \\
\end{array}
$$

Figure 1-5. Major elements in a typical learning trial and the time intervals that separate them.

The extent to which delays influence learning depends on the locus of the delay and the nature of the learning task. Following sections provide a summary of some of the main findings from delay investigations. (Note that the following sections consider learning that is primarily verbal; there is a long history of the study of delays in motor learning that is not considered here.)

Preresponse Delays

Programmed Instruction

In a study investigating the influence of imposed delays between a stimulus and response (Dwyer, Taylor, Canelos, Belland, & Baker, 1985), undergraduate students completed programmed materials on a computer during a single instructional session. There were 57 instructional segments (of one or two frames)
and the stimulus materials consisted of visual and verbal displays of the parts and operation of the human heart (e.g., a diagram showing the heart chambers, with labels of the important features of the display, and an accompanying verbal description). After some learning material had been displayed for a certain amount of time (which varied with the experimental condition) it was removed from the screen. A subject then pressed the return key to receive a question (there were 57 questions in total). Both correct and incorrect responses initiated simple feedback (concerning correctness of the response), but more elaborate feedback (a repeat of the instructional display) was available upon request.

Subjects were paced under one of three sets of conditions: (a) each subject completely determined the rate at which the sequence of events occurred (Self Pacing), (b) stimulus materials were displayed for a predetermined amount of reading time (i.e., 1 s for each text line plus 1 s) plus an additional amount (7 s) of “thinking time”, but were then removed from view (External Pacing with Cognitive Processing Time), and (c) materials were displayed for the predetermined amount of reading time but no thinking time was allowed (External Pacing without Cognitive Processing Time). Hence, the two externally-paced conditions differed in duration of stimulus exposure (i.e., response delay). During programmed delays (i.e., predetermined reading time and think time) learners were “locked out”. That is, key pressing had no effect on the display, pace, sequence, or other aspect of the program.

Five tests, varying in difficulty from recalling names of heart parts to drawing, labelling, and explaining the operation of the heart, followed the learning phase. In each test, learners who studied with externally-imposed delays for reading and thinking (External Pacing with Cognitive Processing Time) performed better than learners who controlled their own pace (Self Pacing), who performed better than learners who studied with externally-imposed delays for reading (External Pacing without Cognitive Processing Time). There are two important findings here: (a) the self-pacing students were poor at adopting the optimal pace for effective learning; and (b) compared to self pacing, external pacing improved performance when it involved longer (External Pacing with Cognitive Processing Time), but not shorter (External Pacing without Cognitive Processing Time) prereaction delays. The additional delay or thinking time made the difference between the delay intervention being more or less effective than self pacing. In addition, the average times required to complete all questions of the program were approximately equivalent for the two externally-paced groups, which were quicker
than the self-paced group. This finding demonstrates that delaying learner responses is not necessarily less efficient than self pacing.

Hence, performance and efficiency were improved by a procedure that delays responding by increasing stimulus exposure time, and provides an opportunity for task-relevant behaviour(s) such as thinking, or inspecting the available materials more closely.

There have been two relevant follow-up studies of Dwyer et al.’s work. One study (Canclos et al., 1985) replicated the consistent advantage of External Pacing with Cognitive Processing Time over Self Pacing at each level of test difficulty. The other study (Kline, 1992) attempted to assess whether learning could be further improved by additional cognitive processing time beyond the 7 s used previously. It tested the notion that if some delay is good, then more delay may be better. The results were not informative; the main outcomes of the previous studies were not replicated, and there were no differences between 0 s, 7 s, 14 s, and unlimited cognitive processing time. However, the discrepant findings may have been due to important methodological differences across studies (e.g., IQ, reading ability, and entering knowledge of subjects; use of embedded questions, and repeated items), so it remains unclear what effect increased stimulus-response delays would have in this implementation of programmed instruction.

Multiple-Choice Tests

It has been shown that performance on multiple-choice computer tests can be improved if no response is permitted for the first 30 s after a question is displayed (Stokes, Halcomb, & Slovacek, 1988). In their study, Stokes et al. had 56 undergraduate psychology students complete multiple-choice tests as part of their course requirements. There were 12 possible tests, each of 10 items, and each test assessed material from a chapter in the course text. Once a question and the alternative answers had been presented on the screen, a programmed delay of 0 s, 30 s, or 60 s was implemented (each subject experienced only one delay condition). The question and multiple choice answers were available during the delay. Because keyboard responses had no programmed consequences during this delay, the student was considered to be “locked out”.

Compared with 60-s and 0-s delay groups, the 30-s delay group not only had higher quiz scores, but also required fewer attempts to achieve this level of performance, and completed the course more quickly. These findings resemble those obtained in the programmed instruction studies; delays improved performance and efficiency. In addition, the 30-s delay group had smaller response-to-question latencies than the 0-s group, indicating that delays increase attentiveness. Although
long delays may result in distraction from the task, when a delay of moderate length (e.g., 30 s) is programmed between presentation of a question and when a response is permitted, the student is likely to read the question (i.e., examine the stimulus features) more slowly and carefully, and thereby make more correct responses.

**Drill and Practice**

The previous studies have demonstrated that performance can be improved by preventing subjects from responding immediately (i.e., by forcing them to wait). A reasonable explanation of these findings was that the delay provided an opportunity to study or think about the materials. In the study described next, a minimum lockout period was not imposed, instead, the existing response period was extended to promote additional study.

For some time it has been believed that the longer a student takes to respond the more likely it is that their response will be incorrect (e.g., Dick, 1965). Computers enable precise control over the timing of instructional events, and some CBI designers limit the amount of time available to make a response in order to reduce procrastination, increase concentration, and consequently increase time on task. However, there may be negative effects of limiting response time. Hativa, Sarig, and Lesgold (1991) maintain that limits on response time may not be beneficial because they induce feelings of pressure, engender opinions that the computer does not allow students enough time to think, and cause the students to make more errors because they guess. Speculating that intervals between a question and response that are too brief may cause children to avoid more reflective strategies, Hativa et al. assessed the consequences of increasing allowable response times for elementary school children completing computer-based practice in arithmetic.

Under normal conditions (i.e., the control group), the maximum time allowed for entering each digit of an answer was 20 s - 30 s and the time limit was 60 s for word problems. The experimental group differed because these time limits were increased to 40 s - 60 s for each digit entered and 180 s for word problems (which Hativa et al. propose is almost equivalent to an absence of time limits). In this way, delays were imposed, but there was no minimum lockout period as there was in the studies described previously.

The increased-time-limit condition showed a reduction in the number of problems attempted in a session (probably because students spent more time thinking before answering), an increase in the number of correct responses, and significantly fewer “time errors” (errors resulting from the student not responding
before the end of the time limit). In fact, time errors were reduced by a factor of 6. Not only did extending permissible response time improve student performance overall, but extending response times was found to be more beneficial for low achievers and students who tended to make large numbers of time errors.

The pacing techniques of imposing prereponse delays (Dwyer et al., 1985; Stokes et al., 1988) and extending response time limits (Hativa et al., 1991) are quite different in their implementation, but they have a similar function in attempting to provide adequate time to respond correctly. Imposing delays improves performance by stopping rapid responding and promoting attention to the materials. Extending available response limits also promotes performance by not cutting off students before they have had enough time to provide a correct answer. The techniques are unlikely to be universally effective, however, and how much time is necessary to perform well, and whether this time should be in the form of a lockout, is likely to depend upon the characteristics of the task (e.g., easy, difficult, new material, or test of familiar material), and subject (e.g., high vs. low achievers, tendency to respond quickly or persist at the task).

Impulsivity

Imposed delays have also been employed to overcome learning difficulties associated with impulsivity. Impulsive behaviour is characterised by quick and inaccurate responding in problem-solving situations, and although it does occur in adults, it is most often a childhood problem (Baer & Neitzel, 1991).

Academic and social behaviours characteristic of impulsive children include poor selective attention, deficient search-and-scan strategies, poor problem-solving strategies, failure to examine response alternatives closely, decreased anxiety about poor intellectual performance, externalised behaviour problems, and aggressive social behaviour (Messer, 1976; Thompson, Teare, & Elliott, 1983). However, whether impulsive responses in academic tasks are part of the same construct as impulsive social behaviour is unclear (Baer & Nietzel, 1991).

For obvious reasons, impulsive behaviour is generally considered to be counterproductive to learning (Egeland, 1974; Thompson et al., 1983). To assist impulsive responders, a variety of treatment approaches have been employed, including self-statement modification (e.g., “go slow, be careful”; Dush, Hirt, & Schroeder, 1989; Meichenbaum & Goodman, 1971), operant procedures (e.g., reinforcement, response cost; Cole & Hartley, 1978), modelling of systematic scanning of stimuli and delayed responding (e.g., Cohen & Przybycien, 1974), and more specific techniques such as teaching children how to scan the stimuli, and how to wait before responding (e.g., Egeland, 1974).
In a meta-analysis that summarised experimental comparisons of each of these treatments with untreated control conditions, the mean ES of treatment was 0.56 (median = 0.47), and there were no reliable statistically significant differences between the effectiveness of different treatment approaches (Baer & Nietzel, 1991). Delayed responding was grouped with other specific strategies for the analysis, and as a group these strategies resulted in a mean ES of 0.68. Effects of this size are relatively easy to comprehend when they refer to changes in academic outcome measures such as examination achievement where norms and distributions are reasonably well established. But how substantial changes of this size are in clinical settings is difficult to interpret.

The strategy of imposing a delay between a stimulus and response is of most interest in this thesis, and its use in an applied setting is demonstrated in the following examples. In one study of the response-delay technique (Dyer, Christian, & Luce, 1982), three autistic children were required to learn one of the following discriminations: (a) identifying whether a presented wooden figure was male or female, (b) raising the hand appropriate to the experimenter’s request, or (c) describing the function of two common objects (e.g., a fork). For each subject, the target behaviour differed according to their ability. In addition to feedback contingent on response correctness (e.g., praise, reprimand, confirmation, primary reinforcers), each subject experienced two conditions: a no-response-delay condition and a response-delay condition. In the no-response-delay condition, subjects could respond immediately after instructions were delivered. In the response-delay condition, a delay prior to responding was accomplished in two subjects by physically restraining them during stimulus presentation, and then for an additional 3-5 seconds before allowing a response. A third subject was able to delay responding without external restraint. Results showed that imposing a delay between stimulus and response increased the percentage of correct responses for each subject. In addition, teachers employed in the research setting who used this technique in their classrooms responded to questionnaires by rating delayed responding as more effective than no delay, and preferable (but not difficult) to implement.

The authors suggested that response delay provides subjects with the opportunity to orient to the relevant stimulus cues before responding, or provides the opportunity to consider alternative solution hypotheses. The latter possibility gains some support from the authors’ report that one child verbalised response alternatives during the delay. Unfortunately for these alternatives, during pilot
research, when delay duration was increased to 6 s, performance was reduced. For subjects on this task, a moderate delay was better than a shorter or longer delay.

In a related study, intellectually-disabled children had improved two-choice colour discriminations when brief delays were imposed between presentation of a stimulus and opportunity to respond (Lowry & Ross, 1975). The four subjects were classified as impulsive on the basis of their quick and inaccurate responding on a visual matching task. During the colour discrimination task, the instructor presented students with a display board with two coloured squares and asked them to touch one of the colours (e.g., “Mike, touch red”). Correct responses were consequtued with M & M’s and verbal praise, whereas incorrect responses resulted in the investigator modelling the correct response or physically guiding subjects in the correct response. The delay condition was implemented by holding the display board out of reach of subjects during the verbal request and for an additional 5 s following the cue; it was then brought within reach to enable responding. In the no-delay condition, the coloured squares were within reach when the verbal cue was given. Each of the four subjects performed at chance levels under no-delay conditions, but when responses were delayed for 5 s, accuracy steadily increased to high levels (approximately 90% correct), and continued at high levels while delay conditions were in effect.

Although imposition of a delay between stimulus and response led to dramatic improvements in accuracy for all subjects, it is unclear exactly why this occurred. The authors suggest that the success of an enforced delay will only be effective if the learner possesses the required process or strategy (e.g., systematic scanning of stimulus and comparison elements) to deal with the task. Hence, if imposed delays are effective when no such skills have been specifically trained, then delays must (a) improve the opportunity for learning such skills, or (b) provide the opportunity to use preexisting skills that have previously been prevented or interfered with by impulsive responses. It is relevant that in neither of the studies detailed were subjects trained in specific strategies for successful task completion. Merely providing a delay was sufficient to gain large increases in performance.

Although imposed delays can improve performance they are not reliably effective. It has been found that delays can increase the latency of responses without decreasing errors (e.g., Baer & Nietzel, 1991; Kagan, Pearson, & Welch, 1966; Thompson et al., 1983). It also has been found that when the specific strategies of delaying responses and scanning systematically are directly compared, each strategy can decrease errors in a visual discrimination task, but only the group
trained in scanning strategies maintained low levels of error in delayed posttests (Egeland, 1974). In fact, teaching impulsive subjects scanning strategies has been a consistently successful means of improving their performance, whereas imposing delays and reinforcing delayed responses have been less reliable interventions (Messer, 1976). Imposed delays do provide an important opportunity for task-relevant behaviour, but are only likely to be effective when appropriately combined with subject skills and task requirements.

**Time Delay**

Like the technique of imposed delays with impulsive responders, time delay involves imposing a delay prior to responding. Additionally, however, time delay includes a prompt for the correct answer which is faded as trials continue. The general procedure is straightforward: a task (e.g., “spell cat”) or stimulus is presented to a learner along with a prompt for the correct response (e.g., student is presented with a card with the word to be spelled written on it), and, initially, the subject needs only to copy the prompt to receive reinforcement for responding correctly. In later trials, however, control over the response is shifted from the prompt to the target stimulus by delaying the prompt for several seconds thereby providing an opportunity for the subject to respond spontaneously (i.e., before being prompted). When subjects respond correctly in the presence of the target without waiting for the prompt, they demonstrate a transfer of stimulus control. Although the use of prompts assists the student when no response is forthcoming, the delay period can be considered important because it promotes examination of stimulus features before a response is made.

Touchette (1971) first used delayed prompts as a means of determining the precise point of transfer of stimulus control. Three severely mentally retarded males were first taught to discriminate red and white coloured keys (illumination and colour were obtained by using projectors and colour filters), which then operated as controlling stimuli in teaching the discrimination of two novel figures: the letter E with legs pointing down and up. The figures were superimposed on the keys using a projector, and colour onset could be used as a prompt signalling the correct figure. By delaying this prompt, the subject could select a figure without assistance; when this happened, the point of transfer of stimulus control from the prompt (colour) to the new stimulus (figure) was identified.

Since Touchette, numerous studies have used delayed prompts to teach new behaviours rather than to measure the point of transfer. The procedure has become known as time delay, and the term actually encompasses a collection of closely-related procedures (see Handen & Zane, 1987). Time delay has been found to be
very effective in teaching nonnormal students (variously described as mentally retarded, autistic, impulsive, and learning or behaviourally disabled) a variety of academic and manual tasks, including spelling (Stevens, Blackhurst, & Slaton, 1991), multiplication facts (Mattingly & Bott, 1990), spontaneous speech (Charlop & Trasowech, 1991; Ingenmey & Van Houten, 1991), function of common objects (Dyer et al., 1982), laundry skills (Miller & Test, 1989) requesting food (Halle, Marshall, & Spradlin, 1979), and food preparation (Schuster, Gast, Wolery, & Guiltinan, 1988) while keeping response errors close to zero (Handen & Zane, 1987; Stevens & Schuster, 1988).

For example, Ault, Gast, and Wolery (1988) used a time-delay procedure to teach mentally-retarded children to read words commonly found on community signs (e.g., exit, danger). The general training procedure involved presenting to the learner a card with the target word printed on it, and asking the learner to pronounce the word. The controlling prompt was modelling of the word by the investigator. During Session 1, after the card and request had been presented, the prompt was presented immediately (0-s delay) so that the likelihood of an incorrect response by the learner was minimal. In subsequent sessions, however, the prompt was delayed by 5 s. This allowed learners to perform the target response before the prompt was provided (known as an anticipatory response). Almost without exception, children performed at 0% level of correctness during baseline probes but, during training with time delay, all words were learned to a criterion of 100% correct anticipations (respond before prompt) on three consecutive trials with a mean error rate of only 0.5% across subjects. High levels of performance continued in maintenance assessments of up to nine weeks, and generalisation occurred across settings and persons: That is, although subjects were instructed in their classroom by their special education teacher, they also performed well in the school library in sessions conducted by the teacher’s aide.

Imposed delays do not guarantee successful anticipations, however, and procedural modifications are sometimes employed to ensure effectiveness (Wolery et al., 1992). For example, Oppenheimer, Saunders, and Spradlin (1993) recently replicated Touchette’s original investigation of delayed prompts with 34 subjects, and found that subjects often waited for the delayed prompt instead of responding prior to it. Failure of time delay to produce correct anticipations was also reported by Glat, Gould, Stoddard, and Sidman (1994) who observed that, during the delay, their subject ignored the response options (alternatives displayed on a card) and looked away. They suggested that the prompt serves as conditioned reinforcer and, because it follows imposed delays, may promote waiting through the delay.
and decrease the likelihood of anticipations. Glat et al. overcame this by requiring their subject to repeat the name of the stimulus (e.g., “bat”) before selecting from the response alternatives. Repeating the stimulus markedly increased correct anticipations, and once performance had met criterion, the repeating requirement could be discontinued with either no reduction in performance or only a small, temporary reduction. The investigators observed that repeating the stimulus set the occasion for scanning the comparisons during the delay. When delay alone was used, this did not happen. These findings show that imposed delays do not ensure correct anticipations, especially when competing behaviours (e.g., waiting for the prompt) are reinforced, but that they do provide the opportunity for precursory behaviours (e.g., scanning response alternatives) that contribute to successful responding once subjects begin to use these precursory behaviours. Other studies have also successfully supplemented time delay with attentional cues (Wolley et al., 1992), and this approach can be likened to supplementing imposed delays with scanning strategies with impulsive responders.

Preresponse Delays: Summary

It has been shown that imposing a brief delay between a stimulus and response, or extending the amount of time available to make a response, can decrease the number of incorrect responses made. This occurred in a diverse range of tasks including programmed instruction, multiple-choice testing, and naming sight words. In each evaluation, imposing a small amount of external control over pacing by inserting delays resulted in better performance than complete self-pacing. Delays appear to provide an opportunity to investigate stimulus features carefully and consider alternative responses before responding. For some subjects it appears that these strategies are not used spontaneously, and must be taught before delaying responses will be advantageous. The following section considers the effects on performance of imposing delays between a response and feedback.

Prefeedback Delays

The previous section showed that, on various tasks, experimentally-imposed delays between a stimulus and response led to better performance than complete self-pacing. A second location within the learning sequence where delays can be imposed is between a response and feedback (i.e., a prefeedback delay).

It is well established that imposing a delay between a response and reinforcer will degrade performance in laboratory animals (Renner, 1964). Most studies with humans have used a quasi-analogue of this procedure by delaying feedback, in the form of knowledge of results, in simple motor tasks. Delayed knowledge of results does not impair immediate performance in such tasks (Adams, 1987; Bilodeau &
Bilodeau, 1958; Bilodeau & Ryan, 1960; Lorge & Thorndike, 1935; Renner, 1964; Salmoni, Schmidt, & Walter, 1984), presumably because humans can employ some verbal mediation or proprioceptive trace to bridge the gap (Boultor, 1964; Renner, 1964).

Studies of feedback delays in human verbal learning have resulted in less uniform findings. In a meta-analysis, J. A. Kulik and Kulik (1988) divided 53 investigations of prefeedback delays in verbal learning into three categories: (a) applied studies with classroom quizzes and programmed materials, (b) list learning experiments, and (c) investigations of acquisition of test content.

**Applied Studies**

An example of an applied study might involve a class of students completing programmed materials. One group of students would receive immediate feedback for each response, and remaining students would have feedback delayed for several seconds. The effectiveness of each condition would be assessed from each group's achievement on a classroom test. Findings from the meta-analysis indicate that immediate feedback typically led to better performance than delayed feedback (mean ES = 0.28). The researchers concluded that, although there may be problems with the internal validity of these studies (for example, sometimes delay of feedback is confounded with amount of feedback), for practical purposes, immediate feedback should be employed in the classroom. This conclusion may be premature. As explained earlier, an ES can be a poor indicator of a typical result and misrepresent the common outcome. A closer consideration of the outcomes from the primary applied studies shows that, of the 11 studies contributing to the mean ES, 4 found statistically significant advantages for immediate feedback, but 7 studies found no differences between immediate and delayed feedback. Even though each of the significant findings supported the use of immediate feedback, the large number of nonsignificant findings mean that it is unclear whether delays are reliably detrimental to performance.

**List Learning**

Studies of list learning included paired-associates tasks, verbal mazes, and stimulus-discrimination tasks. In a paired-associates task for example, learners presented with a stimulus nonsense word, would be required to report or select the correct associate nonsense word. The mean ES in the 27 studies comparing immediate and delayed feedback in this way was 0.34 (i.e., a shift from the 50th to the 62nd percentile), indicating that immediate feedback does, on average, lead to better performance than delayed feedback. Again, consideration of the original outcomes shows that only 10 of the 27 studies found a significant advantage for
immediate feedback, 13 of the 27 studies found no significant differences between
groups, and 4 studies found a significant advantage for delayed feedback. Although
the mean ES indicates that immediate feedback is different from (and better than)
delayed feedback, only half the experiments had statistically significant findings
and some of these favoured delayed feedback. Neither immediate feedback nor
delayed feedback reliably improved performance in list learning experiments.

Informativeness of feedback. Before considering list learning studies further,
it is important to consider the types of feedback that can be presented to subjects.
This can influence the findings of studies employing both prefeedback and
postfeedback delays, and is, therefore, relevant to subsequent sections of this
thesis. Timing is not the only aspect of feedback that can influence performance; the
amount of information contained in feedback is also very important, and some
procedures and terminology require elaboration. Feedback in verbally-based tasks
can be as limited as informing the subject that they were correct or incorrect (e.g.,
by presenting the words “correct” or “incorrect”, or by brief presentation of a
light), or it can be elaborated by providing the correct response by itself (e.g.,
“reflexes”), or both the stimulus and the correct response (e.g., “A doctor taps your
knee (patellar tendon) with a rubber hammer to test your _____”, and “reflexes”).
In some instances explanation of why an answer is correct or incorrect also is
provided. Researchers have referred to the “quality” of feedback to indicate that
some forms are more useful than others. Also, terms such as “exhaustive”,
“complete”, “elaborate”, and “informative” are used to indicate the amount of
information contained in feedback. Because the informativeness of feedback (and
its appropriateness to the task requirements) is very important, and because tasks
vary greatly in the type of feedback employed, following sections will clearly
describe the type of feedback used in each task. The term “complete” will be used
to describe feedback that includes both the stimulus and the correct answer, so that
it provides all the necessary learning information for that trial.

Further analysis of the list learning studies (J. A. Kulik & Kulik, 1988)
indicated that study outcomes also were determined by the informativeness of
feedback. When feedback consisted of the stimulus and correct response, then
delaying this information had little effect on learning (e.g., number of trials to
criterion), but when feedback was the correct response only, delaying this
information reduced learning dramatically. When feedback includes the stimulus
and response it can be considered to be fully informative (i.e., it contains all the
information necessary for learning the association), so a delay before this type of
feedback was ineffective, probably because the stimulus and response are presented
without a temporal separation. When feedback consisted of the response only, then a prefeedback delay did temporally separate the stimulus and response, and performance declined. These findings are analogous to those of investigations of learning in laboratory animals which find that imposing delays prior to feedback (reinforcement) impairs learning.

Acquisition of Test Content

A third type of study of delayed feedback categorised by J. A. Kulik & Kulik (1988) constitutes what has become known as the multiple-choice testing paradigm. In these studies, learners who have done no relevant prior study are required to complete a set of previously unseen multiple-choice test items. Often subjects have no background knowledge of the topic (e.g., Rankin and Trepper's, 1978, subjects who were tested on the biology of human sexuality). Test-item stems are the stimuli and the correct answer is the response to be learned. The tests typically consist of many items (e.g., 60), and, for each item, subjects select one of four multiple-choice options. Feedback may be presented with each item, or it may be delivered after all items have been completed. In the former case, an externally-imposed delay would occur for each item and precede each delivery of feedback; in the latter case, one delay would occur prior to presentation of feedback for all items. Following the first completion of the items, a second (immediate retention) and third (delayed retention) test are completed. Outcomes have been unusual but consistent: Compared to an immediate-feedback group, a delayed-feedback group performed similarly on immediate learning, but subsequently demonstrated improved retention. This has become known as the Delay Retention Effect (DRE).

Using this, or a similar procedure, the DRE has been replicated in many studies (English & Kinzer, 1966; Kulhavy & Anderson, 1972; More, 1969; Rankin & Trepper, 1978; Sassenrath & Yonge, 1968, 1969; Sturges, 1969, 1972, 1978; Sturges, Sarafino, & Donaldson, 1968; Surber & Anderson, 1975). Typically, the retention advantage for the delayed-feedback group over the immediate-feedback group is small, but it appears to be replicable. In their meta-analysis J. A. Kulik and Kulik (1988) report that prefeedback delays improve immediate test performance by an average of 0.36 standard deviations (i.e., more than one grade), and follow-up test performance by an average of 0.44 standard deviations. This effect has occurred with enough consistency to have led some reviewers to agree that delayed feedback is reliably superior to immediate feedback when using multiple-choice tests (e.g., Kulhavy, 1977; J. A. Kulik & Kulik, 1988).

Once again, however, the findings from original studies do not necessarily support average outcomes. In J. A. Kulik and Kulik's meta-analysis, only 7 of 14
individual studies actually reported a significant difference due to delay. Half the studies found no DRE. (The outcomes from follow-up tests in individual studies were not reported.) Because it is not difficult to find experiments that have found no evidence of a DRE (e.g., Cates, 1988; Gaynor, 1981; Newman, Williams, & Hiller, 1974; Peeck, van den Bosch, & Kreupeling, 1985; Phye & Baller, 1970; Sassenrath, 1972; Sassenrath & Spartz, 1972; Sassenrath, Yonge, & Schable, 1968), it is difficult to conclude confidently that prefeedback delays do increase retention.

The DRE is an unusual finding and a theoretically awkward notion; it suggests that although prefeedback delays have no effect at the time of learning, they have a latent effect detectable only at subsequent testing. Nevertheless, from a practical perspective, the DRE is an important finding because it suggests a means of improving long-term retention. Indeed, some researchers have, on the basis of the DRE, suggested that delaying feedback to students would assist them (e.g., More, 1969). This would seem premature considering that many investigations have not actually found the effect. Because the DRE is an interesting finding that has led to a better understanding of how task type, delays, and feedback interact in investigations of human learning, it will now be considered in more detail.

**Brackbill's Studies**

Prior to the multiple-choice test paradigm, the DRE had been demonstrated many times by Brackbill and her colleagues (e.g., Brackbill & Kappy, 1962) using a simple discrimination task. In a typical study, subjects (often elementary school children) learned a series of discriminations. Once the experimental stimuli (e.g., a pair of line drawings of familiar objects) had been presented, a subject selected one. Because the correct stimulus in the pair was arbitrary, the correctness of a subject's first response was determined by chance. Following the delay interval (typically 0 s - 10 s) the light above the correct stimulus would come on, a buzzer would sound, and, if the response was correct, subjects would receive a marble (marbles were conditioned reinforcers backed up by a desired toy). If a response was incorrect, the light above the correct stimulus would come on, a click would sound, but no marble would be presented.

In a series of studies using this experimental paradigm Brackbill and her associates (Brackbill, 1964; Brackbill, Adams, & Reaney, 1967; Brackbill, Boblitt, Davlin, & Wagner, 1963; Brackbill, Bravos, & Starr, 1962; Brackbill, Isaacs, & Smelkinson, 1962; Brackbill & Kappy, 1962; Brackbill, Wagner, & Wilson, 1964; Lintz & Brackbill, 1966) established that delayed feedback during learning did not impair acquisition performance (i.e., the immediate and delayed feedback groups
required the same number of trials to reach criterion, and would make a similar number of errors), but it did improve retention (operationalised as relearning efficiency or savings from acquisition to relearning). This effect of delayed feedback was less reliable at 8-day retention assessments than in retention assessments made one day after acquisition (Brackbill et al., 1967).

**Explanation of the DRE**

**Verbal rehearsal.** To explain the DRE, Brackbill and her colleagues (Brackbill & Kappy, 1962; Brackbill, et al., 1962) proposed a verbal rehearsal mechanism (see also Sassenrath, 1975; Sassenrath & Yonge, 1968, 1969). According to this mechanism, in contrast to a group without prefeedback delays, a group with prefeedback delays has an interval in which they could covertly rehearse material and could practice attending to stimuli; hence, a greater number and variety of criterion-relevant responses will subsequently be reinforced by feedback. If subjects employ such cues, then

(a) the potentially deleterious effects of delay on learning efficiency will be reduced by virtue of a bridging or mediating effect from criterional response to reinforcement, and (b) retention, or resistance to extinction, will be enhanced in proportion to the extent that distinctive response-produced cues have been utilised during acquisition. (Brackbill & Kappy, 1962, p. 17)

The verbal rehearsal explanation was not supported in an empirical test however (Brackbill, 1964). This study is discussed briefly here, but will be discussed in more detail in the next section. When subjects in a delayed-feedback condition were required to complete an interference task (copying pairs of random numbers) during the delay, their acquisition performance was unaffected. This was not expected. On the other hand, the immediate-feedback group, which completed the interference task after receiving feedback, showed impaired performance by requiring more trials to reach criterion and making more errors as they did so. This also was unexpected. Learning was more influenced by events following feedback (postfeedback interval) than events prior to feedback (prefeedback interval).

The verbal rehearsal explanation of the DRE assumed that feedback would reinforce task-relevant activity that occurred during the delay. Brackbill (1964) did not provide a good test of this because she confounded two forms of feedback: By employing visual, auditory, and tangible components in her feedback, it was unclear whether delayed knowledge of results (light and buzzer), delayed reinforcement (marble), or both, determined the results. In an experiment designed to separate these influences, Markowitz and Renner (1966) employed methods
closely resembling those used by Brackbill (i.e., elementary school children, two-choice discrimination task, line drawings), and presented groups with one of several combinations of immediate and delayed reinforcement, immediate and delayed knowledge of results, or no knowledge of results. They demonstrated that the DRE occurred as a result of delayed knowledge of results (i.e., feedback of the correct response) during acquisition, regardless of any delay in reinforcement.

Interference-perseveration. A widely accepted explanation for the DRE is the interference-perseveration hypothesis (Kulhavy & Anderson, 1972). Unlike the verbal rehearsal theory which proposed that prefeedback delay is actively used for rehearsal, the interference-perseveration hypothesis maintains that prefeedback delay has a passive role: During this interval initially incorrect responses are forgotten, so they cannot interfere (proactively) with the acquisition of correct responses obtained from feedback. Contrary to this, subjects experiencing immediate feedback have not forgotten their incorrect responses by the time feedback is delivered (this is known as error perseveration), and initially incorrect responses interfere with acquisition of new material.

Several investigations (e.g., Kulhavy & Anderson, 1972; Sassenrath, 1975; Surber & Anderson, 1975) have provided evidence that supports the assumptions of the interference-perseveration hypothesis. For example, differences have been found between delay and no-delay groups in forgetting and perseveration of errors across tests. Of particular importance is the function of knowledge of results; subjects use this feedback as a means of correcting response errors. Although delays assist the corrective action, it is the presentation of knowledge of results and the subject's subsequent use of it that is fundamental to the process. This also implies that a period immediately following knowledge of results must be employed for active response correction.

In addition to informative feedback, the occurrence of the DRE appears also to depend on a high rate of initial errors which can be corrected using feedback. In Brackbill's discrimination task, and in the multiple-choice-test approach, the material to be learnt was nonsequential and unrelated, and the first response was arbitrary: one study described test item selection in the following way: "the most difficult and obscure anatomical questions were asked in order to reduce the impact of the subjects' previous knowledge of the material" (Rankin & Trepper, 1978, p. 68). Contrary to this, investigations in which subjects studied some material prior to the initial test often do not support the DRE (e.g., English & Kinzer, 1966; Sassenrath, 1972; Sassenrath & Spartz, 1972). It appears that situations resulting
in many errors enable the corrective operation of feedback and, hence, the DRE. Alternately, prior instruction will attenuate the DRE because students will make few errors (Surber & Anderson, 1975).

Although still widely accepted as an explanation of the DRE, the interference-persistence hypothesis has not been uncontested. Some investigators (Peck et al., 1985) have found evidence that challenges the validity of the interference-persistence hypothesis, whereas others (e.g., Sassenrath, 1975) have suggested slightly different mechanisms. For example, Phye and Andre (1989) proposed an attention-based explanation: Prefeedback delay improves retention because subjects attend more closely to feedback after a delay (perhaps because they are less fatigued), and, therefore, they can better utilise the information. Phye and Andre have provided data suggesting that both interference-persistence and attention are important, but that attention is more important.

The likelihood of an investigation obtaining a DRE probably depends on several factors such as the difficulty of the initial task (i.e., the proportion of errors it produces), subject activity in the delays (active rehearsal of stimuli, passive rest to reduce fatigue) and informative feedback. One aspect for which there is agreement is that the primary role of feedback is to help subjects correct their mistakes (Peck et al., 1985). Assuming that feedback is not assimilated instantaneously, the period immediately after feedback should also be important to the correction and learning process.

Prefeedback Delays: Summary

A meta-analysis of applied and list learning studies indicated that imposing delays between response and feedback often impairs learning. In their review, J. A. Kulik and Kulik (1988) concluded that delayed feedback hindered learning. But this conclusion is problematic because many studies have found no difference between groups that are externally-paced with delays and those that are not. Indeed, in some circumstances, delayed feedback may improve retention, as is the case with the DRE.

The DRE was an important finding: Although delayed reinforcement impaired learning in laboratory animals, its analogue, delayed feedback in humans, often had no effect on acquisition but did improve retention. Theoretically, this was important because it advanced the understanding of knowledge of results as something different from reinforcement, but it was of little practical help because of the special experimental conditions required to obtain the effect (repeated exposure to test items that are initially arbitrary and difficult, and for which no previous study has occurred). DRE studies conducted in more regular settings have found no
advantages of employing delays prior to feedback (e.g., Cates, 1988; Newman et al., 1974).

In investigations of prerresponse delays, there was agreement that self pacing was suboptimal. Perhaps this was because learners did not stop to consider the stimulus features or their response alternatives carefully. External pacing by imposing delays provided an opportunity for reflection about the relevant stimulus features and response alternatives. Studies of prefeedback delays have found that the delays are often ineffective, but can be advantageous when used with informative feedback in learning difficult multiple-choice test items. Feedback assists performance because it provides the opportunity for self correction immediately following feedback about the correct response. This indicates that delays following feedback may improve performance by providing an opportunity to use the feedback.

Postfeedback Delays

Misunderstanding of Postfeedback Delays

Theoretical and experimental investigations of externally-imposed delays initially focussed on delayed reinforcement in animals and, subsequently, delayed feedback in humans, so the postfeedback interval received little attention. In fact, early experimental investigations of feedback delays confounded prefeedback delays with postfeedback delays, and there was confusion about the function of verbal feedback, and how it interacted with externally-imposed delay periods. These problems of confounding delay periods, misinterpreting the role of feedback, and overlooking the importance of the postfeedback period to learning is demonstrated in the following examples.

In a study designed to assess the verbal rehearsal explanation of the DRE (Brackbill, 1964; see previous section), subjects had to learn a series of two-choice discriminations under one of two conditions: immediate feedback or delayed feedback. The sequence of events in this study was critical. As Table 1-4 shows, subjects in the immediate-feedback group gave their response, received feedback, completed an interference task (copying pairs of random numbers), waited 10 s, and then began the next trial. Subjects in the delayed-feedback group gave their response, completed the interference task, received feedback, waited 10 s, and then began the next trial.
Table 1-4
Sequence of Events in Brackbill (1964)

<table>
<thead>
<tr>
<th>Immediate feedback</th>
<th>Delayed feedback</th>
</tr>
</thead>
<tbody>
<tr>
<td>Make response</td>
<td>Make response</td>
</tr>
<tr>
<td>Receive feedback</td>
<td>Interference task (10 s)</td>
</tr>
<tr>
<td>Interference task (10 s)</td>
<td>Receive feedback</td>
</tr>
<tr>
<td>Wait 10 s</td>
<td>Wait 10 s</td>
</tr>
<tr>
<td>Next trial</td>
<td>Next trial</td>
</tr>
</tbody>
</table>

The critical difference between the conditions was whether interference occurred before or after feedback was presented. But also note that both groups experienced a 10-s postfeedback delay. For the immediate-feedback group, this postfeedback delay was preceded by the interference task (this group could also be called the “postfeedback-interference group”). For the delayed-feedback group, the postfeedback delay was immediately preceded by feedback (this group could also be called the “prefeedback-interference group”). The effect of this crucial difference in the postfeedback period will be revisited, but first Brackbill’s expectations and the results of the experiment will be described.

It was expected that there would either be no difference in the groups’ acquisition performance, or that the delayed-feedback group would perform worse than the immediate-feedback group because their rehearsal of the stimuli was disrupted. Contrary to expectations, the immediate-feedback group (postfeedback interference) performed much worse than the delayed-feedback group (prefeedback interference) requiring over 60% more trials to reach criterion and making over 80% more errors in the process. Brackbill compared the results with those of a previous experiment which was identical in all respects except that it did not employ an interference task (i.e., a type of cross-study control). Across studies, the performance of the delayed-feedback groups was similar, but the immediate-feedback groups differed: including postfeedback interference reduced performance. Because the findings were quite unexpected, a replication of the interference study was conducted, but there was no change in the outcome: When an interference task occurred in the postfeedback interval, acquisition performance was severely disrupted, but the same task had no effect when it immediately preceded feedback.
Brackbill could not account for the importance of the postfeedback period because she assumed that the feedback she employed functioned as a reinforcer, and that it would have an instantaneous effect. Her feedback consisted of a combination of a light to indicate the correct answer, an auditory cue to indicate the response accuracy, and a marble for a correct response. Brackbill failed to recognise the potentially separate functions of tangible consequences and information about response correctness, and confounded them in this study.

In this study, feedback contained all the information necessary for learning the discrimination (the stimulus item was exposed throughout the trial, and brief visual and audio cues indicated the correct/incorrect answer). Hence, it is likely that subjects experiencing prefeedback interference were unaffected because after feedback had been presented they had all the relevant stimulus materials available and a subsequent 10-s interval in which to make the appropriate associations between the stimuli and correct response. Under this explanation, the postfeedback delay is an active period during which the learning process continues using the feedback, learning does not stop with the presentation of feedback. The postfeedback-interference group performed an interfering task when they could have been making use of the feedback; hence, their learning was impaired. In the 10-s delay period, these subjects did have the stimulus available but had probably forgotten the feedback or had it displaced by the interference task.

Brackbill's study demonstrates how a theoretical and experimental focus on the analogy between delayed reinforcement studies of lower animals and studies of prefeedback delays with humans led to a neglect of the interval following feedback and subsequently limited the theoretical understanding of this aspect of human learning.

**Complex Verbally-Based Tasks**

Brackbill's confounding of the prefeedback and postfeedback intervals was not an isolated incident. This also occurred in one of the first investigations of prefeedback delays with humans (Bourne, 1957). In a concept-identification task, undergraduate psychology students were presented with geometric patterns that had seven dimensions: colour (red or green), form (triangle or square), number (one or two), size (large or small), orientation (upright or tilted), horizontal position (left or right of presentation screen), and vertical position (top or bottom of presentation screen). Only two dimensions were relevant to the correct response, however. A subject's task was to recognise which combination of two dimensions was correct (e.g., large square, large triangle, small square, or small triangle).
Subjects were divided into 18 groups: there were 3 levels of task difficulty (1, 3, and 5 irrelevant dimensions), and 6 levels of delay (0 s, 0.5 s, 1 s, 2 s, 4 s, and 8 s). On each trial, each stimulus pattern was presented for as much time as a subject needed to make a response. Responding involved pressing one of four unlabelled keys which corresponded to the four possible combinations of relevant dimensions. (In the initial trials, subjects’ responses were necessarily arbitrary.) The stimulus was removed from view immediately after a response. Following a delay of 0 s to 8 s, feedback was presented by activating a light above the correct key. The experiment was complete when a subject made 32 consecutive correct identifications.

Time to problem solution, trials to solution, and number of errors made in reaching solution, were significantly and highly correlated and showed similar outcomes, so only error outcomes were reported. It was found that as the duration of prefeedback delay increased, the mean number of errors to solution also increased. In addition, more difficult problems (i.e., those with more irrelevant dimensions) increased errors. The findings reported indicate that prefeedback delays impair performance. This simple conclusion was not warranted, however, because the duration of postfeedback delays was a confounding factor. To keep the interval between a response and presentation of the subsequent stimulus constant, postfeedback delays (9 s to 1 s) decreased as prefeedback delays (0 s to 8 s) increased. Hence, the finding that performance declines as prefeedback delay increases can be reinterpreted as evidence that performance declines as postfeedback delay decreases, or, alternatively, as postfeedback delay increases, performance improves.

Bourne (1957) could not determine whether changes in performance on a concept-identification task were a result of prefeedback delay or postfeedback delay. This confound was overcome in a systematic replication (Bourne & Bunderson, 1963) in which task complexity (one and five irrelevant dimensions), prefeedback delay (0 s, 4 s, and 8 s), and postfeedback delay (1 s, 5 s, and 9 s) were combined factorially. They found that prefeedback delay had no effect on performance, but that increases in the postfeedback delay produced a uniform, linear decrease in the mean number of errors. The advantage of a postfeedback delay was greater for more difficult problems.

Using the same basic methodology, and by manipulating delay length and the informativeness of feedback, Bourne and colleagues (Bourne & Bunderson, 1963; Bourne, Goldstein, & Link, 1964; Bourne, Guy, Dodd, & Justesen, 1965) demonstrated that (a) moderate postfeedback delays (i.e., 10 s to 15 s) produced
better performance than longer and shorter delays; (b) having the stimulus available during the delay led to an additional improvement in performance; and (c) at the longest delay intervals, performance did not suffer if the stimulus pattern was available throughout the interval (Bourne et al., 1965).

Both the postfeedback delay and the information available during the delay determined the outcomes. The researchers concluded that moderate increases in the postfeedback delay improved performance because it provides an "opportunity for S to associate the stimulus pattern with the signaled response category, and/or make appropriate inferences about relevant or irrelevant aspects of the stimulus" (Bourne et al., 1965, p. 628). That is, subjects were "processing" or "rehearsing" the stimulus and feedback information (Bourne & Bunderson, 1963). The researchers also concluded that performance declines with longer delays when stimuli are not present because subjects forget a stimulus pattern; when the next trial arises they cannot recall what hypothesis (about relevant and irrelevant dimensions) they were testing. These negative effects of delays could be countered by making task-relevant information available during the interval.

The importance of combining externally-imposed delays with task-relevant information is also apparent in the following variation of the concept-identification task. White and Schmidt (1972) compared the effects of postfeedback delays to delays imposed between a stimulus and response (preresponse delays). In one experiment, a between-groups design was used with preresponse delay (0 s, 10 s, and 15 s) and postfeedback delay (0 s, 10 s, or 15 s) employed as independent factors. In a second experiment, postfeedback delay (0 s, 10 s, and 20 s) was manipulated within subjects, and preresponse intervals (0 s, 10 s, and 20 s) were varied between groups. Stimuli were not visible during the delays. In both experiments the number of trials required to reach solution did not depend upon preresponse delays, but performance did improve as the length of the postfeedback delay increased. Hence, postfeedback delays improve concept identification performance regardless of whether they are employed between or within subjects.

White and Schmidt found no effect of preresponse delays on performance, even though a variety of other studies have (e.g., programmed instruction task used by Dwyer et al., 1985). This can be explained by the requirements of the concept-identification task. During the preresponse delay, the subject can undertake little activity that contributes to finding a solution that has not already been done on previous trials. Any further progress towards solution depends on receipt of feedback (and a brief interval before the next trial) which is necessary to make inferences about the relevant aspects of the stimuli.
Chapter 1

Concept Identification Studies: Summary

A primary finding from concept-identification studies employing delays is that externally-imposed postfeedback delays are better for learning concept-identification than either no delays, or prefeedback delays. Postfeedback delays appear to be successful when: (a) the feedback or other available information (e.g., stimuli) is relevant to the task solution, (b) the delay is of sufficient duration to enable subjects to use the information, and (c) subjects do make use of the information (as it is possible that without appropriate contingencies for good performance, subjects could ignore the opportunity provided by the delays). When subjects utilise this opportunity, they study the available stimuli, the response, and the association between the two, and discriminate between those aspects of the stimuli which are relevant and those that are irrelevant to the solution.

Paired-Associates Learning

It has been shown that externally-imposed delays improve performance when they provide the opportunity for task-relevant behaviour. Concept identification is a complex verbal task in which the delay is most useful when it follows feedback. The paired-associates task is also verbally based and complex, but differs in important ways from the concept-identification task; for this reason it can provide additional information about the way humans can benefit from delays in such tasks.

Jones and Bourne (1964) used a traditional paired-associates task to assess the effects of prefeedback delay and postfeedback delay on learning of associations between pairs of nonsense trigrams in undergraduate students. They found that delays can improve, impair, or have a neutral effect on performance depending on delay duration and the quality of information available during the delay. In one experiment (Experiment 3), prefeedback delay (0 s, 3 s, and 6 s) and postfeedback delay (0 s, 3 s, and 6 s) were factorially combined, and subjects experienced the following sequence of events: a nonsense trigram was presented for 3 s during which subjects were required to respond with the correct associate trigram; an unfilled prefeedback delay was then presented (in which subjects saw neither the stimulus nor the response), followed by the correct stimulus-response pair and, finally, the postfeedback delay (also unfilled). There were eight pairs of trigrams. Trials continued until the subject made one perfect recitation of the list of trigram pairs. Results showed that increases in both prefeedback delay and postfeedback delay were associated with fewer trials to criterion and fewer errors, and that effects of each delay interval were additive.

Performance improvement due to postfeedback delay was expected, but improvement due to prefeedback delay is contrary to those findings for concept-
identification tasks (e.g., Bourne & Bunderson, 1963). The discrepant findings can be resolved by a consideration of the nature of the tasks. In discrimination tasks (e.g., concept identification), the stimuli consist of both relevant and irrelevant features, and the subject cannot determine those that are relevant until after the feedback presentation. In paired-associate tasks, however, the stimulus is reliably useful so that any delay in feedback presentation can be used effectively to rehearse the stimulus.

The importance of informative feedback. Two subsequent experiments (Jones & Bourne, 1964; Experiments 4 and 5) demonstrated that performance on paired-associate tasks is dependent both on study opportunities provided by externally-imposed delays and the quality of information available during the interval. In Experiment 4 the researchers assessed the effect of “filling” the prefeedback delay by overlapping consecutive trials for blocks of zero, two, four, or eight trials. The zero condition involved receipt of stimulus one, response one, prefeedback delay, feedback, postfeedback delay, then stimulus two etc. In the other conditions, the prefeedback delay for one trial was filled by the presentation of, and response to, the subsequent trial (two-trial condition), that trial and the next three trials (4-trial condition), or that trial and the next seven trials (8-trial condition).

Results showed that the four groups (zero-, two-, four-, and eight-trial conditions) did not differ in the mean number of trials, or errors, to criterion, indicating that prefeedback delays did not interfere with learning in this variation of the paired-associates task. This lack of interference by prefeedback delays was attributed to feedback quality. Because subjects received all the information necessary to learn the association after the prefeedback delay (i.e., feedback consisted of both the stimulus and correct response), filling those delays with subsequent stimulus-response trials did not influence learning efficiency. Direct evidence of the role of the quality of the feedback was provided in a systematic replication of Experiment 4 in which feedback consisted of only the correct response (Experiment 5). Under conditions of degraded feedback, performance declined as a function of the number of successive items presented before feedback. The filler activity apparently interfered with subjects’ ability to maintain the stimulus covertly until the arrival of the correct response.

In summary, Experiment 3 showed that prefeedback and postfeedback delays can improve performance when the delays are unfilled and subjects have task-relevant information available. Experiments 4 and 5 provided evidence that prefeedback delays are used to actively rehearse the materials because when the prefeedback interval was filled, performance declined when feedback consisted of
only the correct response, but performance was unaffected when feedback consisted of the stimulus and correct response.

Several years later, in a follow-up study, Jones (1968) provided evidence that learners will use any delay intervals to study any available task-relevant material. Undergraduate students completed a traditional paired-associates task under one of the following combinations of conditions: prefeedback delay (0 s or 6 s), postfeedback delay (0 s or 6 s), feedback mode (stimulus plus correct response, correct response only), and the presence or not of an interpolated activity during prefeedback delay. The important findings were as follows: (a) across all experimental conditions, postfeedback delays decreased both the number of trials to criterion and errors made; (b) there was an interaction between prefeedback delay and feedback mode: When there was no prefeedback delay, mode of feedback did not influence performance, but when there was a prefeedback delay the stimulus-plus-response feedback group outperformed the response-only feedback group; and (c) generally, the presence of an interpolated task interfered with performance.

Jones proposed a general verbal rehearsal mechanism to account for these findings. A postfeedback delay improves performance because it is analogous to a study period which facilitates “response integration and/or associative hookup” (p. 91). After subjects receive feedback, they have a complete stimulus-response pair to study, so any unfilled period after this will allow rehearsal of the pair and, hence, improve performance. The interaction between presence of prefeedback delay and feedback is a function of whether the delay separates the stimulus from the response. When there is no prefeedback delay, stimulus and response are presented in close succession, and performance is unaffected (this is similar to receiving feedback of both the stimulus and correct response); when prefeedback delays are imposed, stimulus and response are not contiguous, and performance is better with informative feedback (stimulus with response) than when feedback is response only. In the latter case, subjects must restrict their activity during the delay to rehearsing the stimulus until feedback is presented. With informative feedback, however, prior delays can be used to rehearse any previous trial. Finally, the notion that subjects actively perform task-related activities during the prefeedback delay is supported by the finding that interpolated tasks during this interval impairs performance.

Jones’ (1968) study is not the only one to have found an effect of the quality of feedback provided during prefeedback delays. As alluded to earlier, J. A. Kulik and Kulik (1988) have reviewed many paired-associates studies that employed prefeedback delays, and supported Jones’ conclusion that, when feedback includes
stimulus and response, then the imposition of a delay before feedback does not impair performance. When delayed feedback does not include the stimulus (i.e., response only), however, performance will be impaired.

Finally, it is helpful to consider the conditions experienced by the best performing group in Jones (1968). Those subjects had elaborate feedback, no interpolated task, and experienced delays both prior to, and following, feedback. If the rehearsal theory is correct, then these subjects outperformed the other groups because they were able to utilise the prefeedback delay to continue rehearsing paired associates from previous trials, and the postfeedback delay and elaborate feedback provided to rehearse the current trial’s associate pair.

In summary, to be successful, external pacing by imposed delays must take into account not only the locus of delay, but also the nature of the task, the informativeness of feedback, and the duration of the delay. Although Jones established that subjects may use any available opportunities to work on the task at hand, the advantage of postfeedback delays is often that all other task-relevant information is available at this point.

**Delays in System Response Time and User Response Time**

A considerable amount of information regarding the effect of delays on human performance is contained in the human-computer interaction literature. Before the development of personal computers, a large amount of work was conducted on time-share, and server-based, systems. When the number of people using these systems increases, system response time decreases sharply (Doherty & Kelisky, 1979). Hence, an important concern to purchasers and users of those interactive systems was how delays in system response influenced work efficiency. Before experimental investigations were conducted it was commonly assumed that delays in computer responsiveness of more than a couple of seconds would degrade user performance (Dannenbring, 1983), but subsequent empirical investigations have shown that system delays may have a detrimental, neutral, or beneficial effect on performance depending on task requirements.

Figure 1–6 shows the temporal sequence of events and delay intervals occurring in interactive computing systems. This is a general representation only, it cannot show all variations of events. In particular, “feedback” and “task” are loosely defined. Sometimes (such as in data entry tasks) feedback may not be presented explicitly, and users merely have control of the machine returned to them so that they can continue with what they were doing. At other times (e.g., when debugging a computer program), however, feedback can be presented as information regarding the execution of the program and, hence, subjects continue
the original task but in a form modified by the feedback. In another variation, in which there are separate tasks in distinct trials, any feedback that is presented is relevant only to the preceding trial, and the next trial is a separate and distinct event. In an ongoing computer-user interaction, there are two types of delays, those due to the system (System Response Time, SRT, defined as the time interval that begins once a command is entered, and terminates as soon as the computer returns control to the user), and those due to the user (User Response Time, URT, defined as the time interval that begins after control is returned to the user and terminates when the user inputs the next command).

![Diagram of SRT and URT](image)

**Figure 1-6.** A typical sequence of events and location of delay intervals occurring in interactive computing systems.

**System Response Delays**

Initial observational data lent support to the notion that delays in SRT are detrimental to task efficiency because they produce a similar delay in URT (Doherty & Kelisky, 1979). Hence, the decrease in efficiency caused by system delays are compounded by resultant user delays (see also Goodman & Spence, 1981). In addition to the negative effects of delay duration, variability in delay length was regarded as especially detrimental, with a long, erratic delay contributing to slower work, emotional upset, and increased errors (Doherty & Kelisky, 1979).

There is some empirical support for these data. For example, in an investigation of the effect of the length and variability of SRT on performance (Butler, 1984), subjects (experienced keyboard users) were presented with a typed list of five-character letter groups and were required to enter them at a keyboard. After typing the last letter of a group, subjects pressed the carriage return which invoked an SRT. After the SRT, a prompt signalled when data entry could recommence. There were 10 SRT conditions, and each subject completed each condition. Five different mean computer response times were used: 2 s, 4 s, 8 s, 16 s, and 32 s, and for each mean value, there was a high and low variability condition because variable SRT's more accurately represent the function of the computer than do constant SRT's (Butler, 1984). For example, the conditions with
a mean SRT of 2 s, the low-variability condition presented delays of approximately 1.6 s to 2.7 s, whereas the high-variability condition presented delays of approximately 0.8 s to 4.2 s. (SRT ranges were based on calculations of just noticeable differences.)

There were two important findings: (a) there were no differences between groups in the number of errors made and typing speed, and (b) URT did increase as a negatively accelerated function of increases in SRT. In addition, high variability SRT increased URT. So although SRT did not affect performance, it did impair efficiency.

This SRT-URT relationship has been explained (Doherty & Kelisky, 1979) with reference to a short-term memory effect: If a system delay is longer than a second, the subject forgets the sequence of keyboard actions, and must reconstruct the sequence which increases URT. Other guidelines for the maximum allowable delay duration proposed that they should not exceed the duration of delays which occur normally in human conversation. Again, this is explained by the disruption of short-term memory (see Dannenbring, 1984).

The logical implication of these findings is that by reducing the duration of system delays, users will not be adversely affected, and performance and efficiency will increase (e.g., Butler, 1984; Doherty & Kelisky, 1979; Goodman & Spence, 1981; Lambert, 1984; Thadhani, 1981). Other investigations, however, have shown that moderate SRT’s do not impair performance, and may even improve it (e.g. Dannenbring, 1983; Grossberg, Wiesen, & Yntema, 1976). The reason for these discrepancies appears to be the level of task complexity: With simple nontasking tasks, increases in SRT impair performance (or efficiency), but with complex and demanding tasks, increases in SRT improve performance.

Simple tasks. One of the first empirical investigations of the effects of SRT (Dannenbring, 1984; Experiment 1) assessed the hypothesis that delays in responsiveness greater than those found in normal human conversation will be disruptive. For 30 min subjects interacted with a simulated psychologist engaging in nondirective therapy. After subjects entered a response, they received one of the following types of delays: 0 s (no delay), 5 s fixed, 5 s variable (ranging from 1 to 9 s), 10 s fixed, and 10 s variable (ranging from 6 to 14 s). The only differences in performance across the delay groups reflected that more typing, or "conversation", is possible with shorter delays. Groups did not differ in their perceptions of such things as the computer’s performance, intelligence, or friendliness, and they equally enjoyed interacting with the system. In a simple
interactive dialogue task, SRT delays longer than those which occur in normal human conversation are inefficient but not annoying or disruptive.

A second experiment (Dannenbring, 1984; Experiment 2) investigated the effect of SRT on a more typical keyboard activity, data entry. Subjects were required to key in 160 three-digit numbers from a typed list (arranged in 40 rows of 4 numbers). The delay durations and procedures of the first study were repeated. Again, the only aspect of subject performance and perception that was affected by delays was time to complete the task: groups experiencing longer delays required more time overall. The investigator suggested that increases in SRT increase task completion time because the task is simple, requires little thought, and, hence, the subject simply waits for the next opportunity to make a response. A more interesting finding, and one with potentially greater implications, arose from the comparison of the groups' time on task (calculated by subtracting the enforced delays from the overall completion times). It was discovered that longer delays were associated with quicker performance. Although it cannot be determined from the data presented, longer SRT's may have reduced the time on task because subjects type faster to compensate for time lost due to enforced delays. In their multiple-choice testing study, Stokes et al. (1988) also found that delays are associated with increased efficiency.

System delays led to user delays, but, in simple tasks, their influence appeared to be mostly on the interval between responses to the machine. The quality of the response (correctness, speed) was either not influenced (Butler, 1984; Dannenbring, 1984, Experiment 1) or was improved (Dannenbring, 1984, Experiment 2).

Complex tasks. One way in which SRT can improve performance is to promote increased attention and carefulness in relatively complex tasks. In one study, programmers were required to debug a program as quickly and accurately as possible within a 30-min period (Dannenbring, 1983). Corrections were made by retyping the entire line which contained the error. Whenever a correction was entered, subjects were immediately presented with one of the following system delays: 0 s (no delay), 5 s fixed, 5 s variable (ranging from 1 s to 9 s), 10 s fixed, and 10 s variable (ranging from 6 s to 14 s). There was no effect of any delay interval on any of the following measures: ratings of the task difficulty, total number of program runs before the correct solution was gained, total number of errors made by subjects, or time spent on task (i.e., total time minus enforced delays). Subjects did, however, make fewer line deletions as delay size increased. That is, the same performance levels were achieved with fewer correction entries. It
was observed that subjects were often working during the enforced delays, and this is evidence that delays provide an opportunity to study the problem and, hence, make more efficient corrections. It is also possible that the presence of a delay was aversive, and that subjects avoided them by being careful and making each entry maximally effective. But this was not supported by ratings of satisfaction with the performance of the computer, which were not different between delay groups.

Similar improvements in carefulness and efficiency have been found in problem solving tasks in which subjects control the occurrence of system delays (e.g., Grossberg et al., 1976). In this study, subjects worked on a system which only replied to commands that requested output, or were in error. Other commands (e.g., those requiring calculations to be performed) were accepted and processed, but no output was provided. Due to this mode of operation, users typically typed a series of commands which ends in an output request. If any errors were made during command entry, the machine would interrupt ongoing typing, display the error message, discard typing subsequent to the error, and await new commands. Delays averaging 1 s, 4 s, 16 s, and 64 s were imposed between a request for output and the output itself. Subjects worked on tasks requiring approximately 15 min to complete, and which were described as involving calculations on arrays of numbers, and requiring a thoughtful and systematic approach.

It was found that as delay length increased, intervals between commands (both request and nonrequest) to the machine increased, intervals between commands that produced output increased even more rapidly, and the total number of commands necessary to perform the task decreased. In addition, illegal commands decreased. These findings provide further support for the notion that, as a result of delays, subjects become more cautious and deliberate, and use their commands efficiently: There were no differences across delay conditions in the time required to reach a solution. In this experiment, delays may have operated as response-contingent aversive stimuli (punishers) which were avoided by not entering illegal or inefficient (minimal gain per request) commands. Unlike the previous study, the advantages of delays in this study did have a cost, as long delays “aroused strong emotions” (p. 222).

SRT is a system-controlled delay in task progression that was uncontrollable and unavoidable in realistic settings, and was assumed to result in performance decrements such as increased errors, slower work rate, and emotional upset. Evidence shows, however, that in simple tasks such delays do little to performance other than reduce the rate of responding, but, on more difficult tasks, can increase attention to the task and provide the opportunity for study and planning. Delays
were not necessarily annoying, although this was found sometimes. The advantages of delays are more readily seen in investigations of the user response time.

**User Response Delays**

URT is commonly known as “think time” (Doherty & Kelisky, 1979). By definition, URT includes the time prior to a response (think time) plus the actual time to make a response, but it is typically assumed that the physical response constitutes a minor portion of the interval (see Figure 1-6). Although URT is typically under the complete control of the user, an experimental technique employed to study the effect of this interval on behaviour consists of “locking” a keyboard for a prescribed period of time. During this “artificial” URT the subject has access to the task, but not the computer.

An example of this procedure was provided in a problem solving task in which, except for a 2-hr limit, the number of trials to solution, and subject access to information, was largely unconstrained (Boehm, Seven, & Watson, 1971). Subjects were presented with a map of an urban area which contained a grid of surface streets and freeways, and which provided information concerning the location and frequency of emergencies. The subject’s task was to find, within the time limit, the optimum location for emergency hospitals. To assist them, subjects had access to a computer which was programmed to accept as input their proposals for hospital location, and return information regarding the effectiveness of the location (i.e., emergency response time). For some subjects the keyboard was locked for 5 min or 8 min after the computer had returned feedback, whereas other subjects had free access to the computer. The 5-min lockout group outperformed the free-access group and the 8-min lockout group by minimising emergency response times within the 2-hour period. In addition, it was found that, of the 5-min lockout group, members less experienced in computing and operations research performed better than their experience would predict and, of the long-delay, and no-delay, groups, the more experienced members performed worse than their experience would suggest. Moderate lockout periods imposed between presentation of feedback about the current solution and opportunity to submit a modified solution, were more effective than either no lockout or longer lockout periods.

One finding of this experiment was that subjects typically expressed dissatisfaction with keyboard lockouts even though these restrictions on access improved performance on the task. The technique improved performance at the expense of user satisfaction. Compared to many delay investigations, however, the
delay sizes employed in this study were very large, and it is possible that delay intervals in between 0 min and 5 min may have increased performance without causing dissatisfaction with the technique.

In summary, studies of human-computer interaction with externally-controlled delays for time-share systems do not find general reductions in performance and emotional effects. On simple tasks such as data entry, machine delays do not reduce performance, but do impair efficiency (e.g., Butler, 1984, Dannenbring, 1984, Experiment 1) because the user waits for the next response opportunity. On more complex tasks (e.g., debugging programs), delays can improve performance (Boehm et al., 1971) and efficiency (Dannenbring, 1983; 1984, Experiment 2; Grossberg et al., 1976) because subjects use the opportunity to work on the task. Although externally-imposed delays may result in negative emotional reactions (e.g., Boehm et al., 1971; Grossberg, et al., 1976), this does not necessarily occur (e.g., Dannenbring, 1983; 1984). Dissatisfaction may be related to delay duration: In the studies reported above, negative reactions were reported in studies using durations of one to several minutes, but not in studies using briefer durations (up to 10 s).

Finally, it is worth mentioning that personal computers have removed SRT's as an issue, because these devices are extremely fast single-user systems which typically respond without noticeable delay. The intelligent use of URT's or lockout procedures appear to have some merit, however, and will be discussed in more detail in the following sections.

Experimentally-Imposed Delays: Summary

Student self pacing has been advocated as a necessary component of individualised and automated instruction since Pressey's multiple-choice testing device and Skinner's programmed instruction, and it continues to be employed in modern CBI. Although self pacing is typically operationalised as allowing a student to progress through materials completely free of external control, there is evidence that external control of pacing, by judicious use of imposed delays, can improve learning.

Performance is enhanced when delays are imposed in a variety of tasks (e.g., demanding computer tasks, and discrimination tasks such as colour and concept identification, and programmed instruction) and in certain populations (e.g., impulsive responders, college students). Delays have clearly been beneficial when imposed between a stimulus and response (prerresponse delay), or between feedback and the next stimulus (postfeedback delay); delays between a response and feedback (prefeedback delay), were not consistently beneficial. The locus of
the delay is less important than the opportunity for task-relevant activity that the delay provides.

For impulsive responders completing match to sample tasks, prerresponse delays provide the opportunity for behaviours such as systematic scanning of stimuli that may previously have been blocked by impulsive responses. For complex verbally-based tasks, prerresponse delays were found to be effective in improving performance because they provide an opportunity to read the information more slowly and carefully, or process information more deeply (e.g., Dwyer et al., 1985; Hativa et al., 1991; Stokes et al., 1988). Delays imposed after feedback have been found consistently to be effective in improving performance because they provide an opportunity for rehearsal (e.g., Bourne & Bunderson, 1963), to formulate new hypotheses about the correct solutions, or to draw appropriate inferences (Bourne et al., 1964; Bourne et al., 1965). Hence, externally-imposed delays provide learners with an enforced opportunity to engage in task-relevant behaviour which will assist learning if the information available during the delay is useful in the problem solution and subjects make use of the information and the opportunity.

**Rationale and the Present Studies**

These experiments assessed the tenet that students perform best when they control their own study pace. Although this tenet is axiomatic in CBI, few researchers have formally tested the notion (Kline, 1992), and those that have (e.g., Cancllos et al., 1985; Dwyer et al., 1985) found that total user control does not optimise performance.

Indeed, the literature review has shown that external control of pacing using imposed delays has improved performance in a variety of tasks (e.g., concept identification, multiple choice tests) and samples (e.g., impulsive responders, college students). Although the precise function of a delay appears to differ depending on the demands of the task and skills of the learner, externally-imposed delays generally provide an opportunity for study that is not otherwise taken. Study may consist of inspecting stimulus features more closely and carefully, considering alternative responses, or overt or covert rehearsal of the study materials. The literature also indicates that there are several important aspects of delays that contribute to their effectiveness. Delays must be located at a place in the stimulus-response-feedback chain where there is suitable information available for study purposes. Generally, prerresponse delays and postfeedback delays improved performance, but prefeedback delays did not. Delays were most useful when they
Chapter 1

were of a moderate length. Short delays do not provide enough opportunity to work on the task, whereas long delays can cause distraction or forgetting. More importantly, longer delays were useful so long as learners had task-relevant information available for study throughout the delay. Also, delays are useful only if the learner has relevant skills to employ during this interval.

In the following studies, external control over pacing was accomplished by imposing brief noncontingent postfeedback delays in the form of a keyboard lockout. Because learners can only be ensured of having all the necessary instructional components of a trial after the problem has been presented, a response has been emitted, and informative feedback has been presented, the most profitable place to employ delays is likely to be following feedback. By employing postfeedback delays in conjunction with complete feedback (i.e., the stimulus, response, and correct answer were all available during the delay), conditions for learning were maximised. Learners were university students and were likely to have good study skills at their disposal.

By using delays in this way, learners maintained control over their rate of progress throughout most of the instructional sequence, but were also given the opportunity to study the materials once feedback had been presented. This was a minimally intrusive procedure that attempted to maintain the affective benefits ascribed to self control, and also provide the performance benefits associated with external control of pacing.

Although pacing was the primary issue in these investigations, there were several other important considerations.

1. Programmed materials. One issue is whether the advantages of delays transfer from less complex tasks such as concept identification and paired associates, to more complex tasks such as realistic instructional materials. The following studies used programmed materials (i.e., Holland & Skinner, 1961) because programming has good theoretical and empirical credentials, and is likely to be a profitable basis for continued research of instructional approaches.

Using delays with well-designed realistic materials is not only a good test of the efficacy of delays, but it represents a research emphasis on instructional approach rather than delivery mode (i.e., the medium). Although fundamental instructional principles have been widely researched, they are not well known, and, to the detriment of courseware quality, they fail to make their way into CBI (Tudor & Bostow, 1991; Vargas & Vargas, 1991). Because an instructional medium (such as a computer) is only as good as the programs and contingencies it presents to students, computers are unlikely to be a revolutionary instructional device until a
technology of learning is incorporated into programs. Hence, to complement the ongoing advancement of digital media, it is necessary to continue researching instructional approaches.

2. Duration of experiment. Many studies of performance with programmed instruction have been inconclusive because they presented only a small number of frames (e.g., Boersma, 1966; Kulhavy & Stock, 1989, Experiments I, II, & III), or used programs of unknown quality (see Holland, 1967; Tobias, 1973). In contrast, the present experiments employed over 1500 frames which covered 40 or more topics of a well-tested text (Holland & Skinner, 1961). This is important. If only a small number of frames is employed in one brief session, it is unlikely that large differences between conditions will emerge. Furthermore, previous studies have generally used group-based analyses on these data, so inter-subject variability may have obscured differences between conditions even further (Crosbie & Kelly, 1994). Because the following studies were relatively long, both the extent and reliability of impact of the experimental variables could be determined.

3. Experimental design. Very few studies of programmed instruction have employed a single-subject design; an element which would be expected in traditional behavioural methodology. A single-subject design was considered to be advantageous in the following studies for several reasons. Some of these are related to the need for a sensitive tool to detect differences across conditions. By using realistic educational materials, extraneous factors (such as difficulty or length of a topic) are likely to be introduced that increase variation in the data (“noise”), and may mask subtle effects of experimental manipulations. The use of a group-based design may further increase noise because individual differences are included in the data for each group. (In addition, because these designs potentially confound inter-subject differences and experimental differences, they have problems with interpreting causal influences.) In addition to problems with noise, high levels of accuracy (as often occurs with programmed materials) may make subtle treatment effects difficult to detect. The considerations involved in choosing the experimental design are discussed fully in the General Method.

4. Delivery medium. It was important to assess the self-pacing tenet in a computer medium for two reasons: (a) computers enable tight control of experimental contingencies, and provide accurate and precise time-based measurements; and (b) computers are ubiquitous in present-day classrooms and their use in teaching is likely to increase in the future, but they may promote suboptimal performance because their speed of operation minimises delays and encourages fast responding.
5. Social validity: Although some studies have found that external pacing annoys subjects (e.g., Boehm et al., 1971), other studies have not (e.g., Dannenbring, 1983), and it is unclear whether satisfaction with these procedures is related to delay duration. Furthermore, because investigations of delays have typically been very brief, it is unclear whether opinions about delays change with long-term exposure to them. Perhaps delays are better tolerated over shorter terms than longer terms. Because the acceptability of behavioural procedures to their consumers is a determinant of the procedure's overall success, the present experiments assessed the social validity of the pacing approaches (Geller, 1991, Social validity; Wolf, 1978).

In light of these considerations, an investigation of the effects of self pacing and external pacing was operationalised using the following general approach: Each subject studied topics from Holland and Skinner's (1961) programmed text each week day for three to four weeks. On each day they studied one topic under self-pacing conditions, a second topic under imposed-delay conditions, and possibly a third topic under other experiment-specific conditions. When working through a topic on the computer, subjects read a small amount of information and a question related to this information, provided an answer, received feedback, corrected their answer, and continued to the next item. Under self-pacing conditions, subjects determined how quickly or slowly they completed these steps. Under imposed delays, the only difference was that subjects could not immediately proceed to the next question after receiving feedback because a brief delay was imposed. During this time the question, subject's response, and correct answer were displayed, the keyboard was locked, and responding had no programmed consequences. In addition to presenting the materials and experimental contingencies, the program recorded information regarding a subject's accuracy and efficiency, and, at the end of each topic, it prompted the subject to rate how satisfied they had been with the experimental session. In some studies, subjects also completed immediate and 1-month delayed posttests to determine relative retention for materials learned under externally-paced and self-paced conditions.
CHAPTER 2: GENERAL METHOD

Experimental Design

Single-Subject Designs

Programmed instruction developed as a technical application of behavioural principles that had been established in the operant laboratory. Single-subject designs have typically been the major tool of operant investigation but they have been used only infrequently in programmed-instruction research. A major aim of the present studies was to use a single-subject methodology with programmed instruction and realistic educational materials (where instructional sets differ in length, content, and difficulty).

Behavioural researchers have characteristically employed single-subject designs for several reasons: (a) correlational procedures (e.g., intact groups designs) make determination of causality impossible (Sidman, 1960); (b) intact groups designs confound inter-subject differences with experimental manipulations (Sidman, 1960), so general conclusions regarding the effect of the manipulation on a specific individual, within or outside the sample, are, therefore, impossible; (c) single-subject designs require fewer subjects which often translates into both time and cost savings; (d) the repeated measures associated with single-subject designs continually assess behaviour over time; (e) any adverse effects of treatment on the behaviour recorded is obvious; and (f) any effects of the treatment(s) are replicated within the same subject over time. Another reason for using single-subject designs in the current experiments was to assist with the examination of performance during the learning of academic materials. By using realistic materials inter-subject variability in performance predictably would be large, but the effect of experimental conditions (e.g., imposed delay) may be subtle and differ across individuals. Anything other than a single-subject design would mask such effects.

Alternating Conditions Designs

Some single-subject designs require the sequential implementation of two or more experimental conditions so that each subject completes the first condition before commencing the next one (e.g., multiple-baseline, changing criterion, reversal). With those kinds of designs any condition effects are confounded with order effects because, at each phase in which a condition is implemented, the subject may have been changed by the previous condition. Those designs, therefore, restrict conclusions about the effect of the intervention (Kazdin, 1982; Kazdin & Hartmann, 1978), and the interpretive difficulties posed by sequence or order effects become greater as the number of conditions to be evaluated increases.
Furthermore, those designs require stable levels of behaviour under one condition before the next condition can be introduced, which may take many sessions, or may not occur at all (this is discussed further shortly).

An ideal experimental design would split a subject into as many identical organisms as there are conditions, so that each organism, and the accompanying experimental variables, could be compared when background variables are identical and when there is no possibility of a sequence effect arising from the procedure (Sidman, 1960). Such a single-subject design would overcome the problems mentioned above by removing both inter-subject and intra-subject variability. Sidman described such a procedure. Behaviour is measured repeatedly under rapidly alternating conditions. That is, rather than using a single sequence of conditions like the designs mentioned above, a sequence of conditions is repeated many times. To do this, each condition has a distinct discriminative stimulus (SD) and alternations occur independent of the level of the subject’s behaviour. Using this procedure, control over behaviour would be demonstrated by a clear difference in responding under each SD. This basic procedure has been called many things: alternating conditions (Ulman & Sulzer-Azaroff, 1975), alternating treatments (Barlow & Hayes, 1979; Barlow, Hayes, & Nelson, 1984), multielement baseline (Sidman, 1960), multiple schedule (Barlow & Hersen, 1973; Ferster & Skinner, 1957; Hersen & Barlow, 1976), randomisation (Edgington, 1967), and simultaneous treatment design (Kazdin & Hartmann, 1978).

Variations of the basic model and differences in terminology are due mainly to differences in the origins of the procedure rather than strict technical differences (although the term “multiple schedule” is normally restricted to designs in which alternate conditions differ only in the schedule of reinforcement used). The term “alternating conditions design” (Ulman & Sulzer-Azaroff, 1975) was proposed as a more descriptive alternative to Sidman’s “multielement baseline”, because it emphasises the experimental manipulation rather than the behavioural effects. This term will be used in the current thesis because of its reference to the experimental technique and its neutrality with respect to historical influences and field-specific implications (see Barlow & Hayes, Note 1).

Although single-subject designs demonstrate numerous advantages over correlation- or group-based alternatives, they are not completely free of interpretive difficulties (Barlow & Hersen, 1984; Kazdin, 1982). It is important to recognise potential problems in single-subject designs and justify the appropriateness of the alternating conditions design for the current set of experiments. Because the reversal design has been a typical evaluative tool in applied research (Kazdin, 1984;
Ulman & Sulzer-Azaroff, 1975), it will be used to illustrate the relative merit of the alternating conditions design (Crosbie, 1987).

**A Reversal Design**

The simplest reversal design has three distinct phases: The behaviour of interest is first measured under baseline conditions (A) until it reaches a stable level. A second phase then commences in which the experimental condition (B) replaces Condition A, and measurements continue to be taken until the level of behaviour stabilises or diverges from the level predicted by baseline. In a third phase (the reversal phase), baseline conditions (A) are reinstated. Differential influence of conditions A and B is then assessed by comparing behaviour between baseline, experimental, and reversal phases. Experimental control over behaviour is demonstrated if the level of behaviour during Phase B is clearly different from that of Phase A, but subsequently reverts to its original level during the reversal phase. In each of these phases, many data points should be gathered to establish the stability or directional trend of the behaviour of interest. If behaviour does not stabilise, or does not reverse, then results can be uninterpretable. These problems are discussed in more detail shortly.

This reversal design (ABA) requires three phases to assess one experimental intervention (a phase of Condition B), and, because the introduction of Condition B coincides with change per se, it is not possible to know which of these factors is responsible for any change in the level of behaviour. To rule out this alternative explanation the design is often extended by adding a fourth measurement phase with an existing (ABAB), or new (ABAC) treatment to further demonstrate experimental control over the behaviour. Further extended designs (e.g., ABACABA) can also be used to assess multiple conditions relative to baseline.

**Alternating Conditions Designs: An Alternative to Reversal Designs**

When experimental conditions are evaluated using an alternating conditions design, successive measures of behaviour are taken while conditions are alternated rapidly. In this way each condition is compared within a single intervention phase (cf. across three phases in a reversal design). The sequence of conditions in a reversal and alternating conditions design is depicted in Figure 2-1 where, in each design, a “column” represents a phase of multiple measures under that (those) condition(s).
Reversal design Alternating conditions design

**Figure 2.1.** Between-phase (left) and within-phase (right) alternations of experimental conditions.

The two designs are closely related because the alternating conditions design is merely a "fast-paced reversal design incorporating many reversals" (Hains & Baer, 1989, p. 57). How fast the alternations occur varies considerably: for example, from as often as every two minutes in the operant laboratory, to one session of each condition per day in clinical studies. A distinct discriminative stimulus accompanies each condition and may range in complexity from a coloured light to a large printed sign. The alternation of conditions can be systematic or random and, importantly, it is independent of subjects’ behaviour. Unlike other single-subject designs (e.g., reversal, multiple-baseline) there is no prerequisite for behaviour to be stable before a change in conditions is applied. Control over behaviour is demonstrated in an alternating conditions design if there is a consistent difference in the level of behaviour between the alternating conditions which should be apparent from visual inspection.

Baseline and reversal phases are commonly included in alternating conditions designs. A baseline phase (Condition A only) is included to show the level of a behaviour under "normal" conditions and provide a reference for which to compare changes during the alternating phase. A reversal phase may be included to further examine the level of behaviour influenced by one condition in isolation. In applied settings particularly, the most effective intervention is commonly used in the final phase to maximise behavioural gains. Baseline and reversal phases may be useful in some applications but are not essential. The most important aspect of an alternating conditions design is that there are many alternations of the experimental conditions; hence, enough experimental comparisons are obtained to demonstrate different levels of control by alternate sets of conditions, and order effects are controlled because AB alternations are as frequent as BA alternations.

**Advantages of an Alternating Conditions Design**

**Background variables.** Concomitant changes in variables such as practice and fatigue are potential threats to relatively slowly changing designs (such as the reversal). The alternating conditions design can provide experimental comparisons in the presence of these background variables. The design presumes that the rate of
change in experimental conditions is faster than that of uncontrolled background variables. This means that condition effects will not be confounded with these extraneous influences.

**Stability.** A reversal design requires that there be obvious changes in the level of behaviour from the baseline phase to the experimental phase. To demonstrate this, normal variability in behaviour ("noise") during baseline must be minimal so that it does not obscure a change in the level of behaviour caused by the experimental manipulation. In addition, baseline behaviour must not show a trend. If it does, and the experimental manipulation was intended to shift behaviour levels in the same direction, then the experimental effect cannot be separated from natural trend as an explanation of the changes in behaviour.

Some behaviours do not readily conform to objectively defined criteria for variability and trend. In academic fields it is not uncommon for baselines to be ascending and for performance to be cyclic (Ulman & Sulzer-Azaroff, 1975), and a large amount of time and many sessions may be wasted if the target behaviour will not stabilise. Hence, the reversal design is not useful in these situations, but the alternating conditions design, which has no stability or trend prerequisites is more appropriate.

**Nonreversal.** Before conclusions about causality can be drawn, a reversal design also requires behaviour to revert to baseline levels when baseline conditions are reintroduced. Without this demonstration of control, interpretive difficulties arise because it is unclear whether changes in the level of behaviour during the experimental phase were due to the treatment or background variables. Failure to reverse can occur if the behaviour has been permanently altered by the intervention such as when natural contingencies take over from experimental ones (e.g., newly acquired grooming or oral hygiene behaviours may be maintained by social reinforcers such as attention), or the experiment is terminated prematurely (before a stable reversal). Premature termination is possibly the worst scenario in a reversal design. Because conclusions are based on behaviour levels across phases, it is essential that all phases are completed to provide a result. Ultimately, reversal designs require the investment of much time and many experimental sessions which may be misspent if stability cannot be achieved and control via reversibility is not demonstrated.

By comparison, in a phase of alternating conditions, valid comparisons of conditions are collected quickly (one comparison for each pair of conditions presented), and the experiment can be terminated as soon as control is
demonstrated. If early termination occurs, sufficient comparisons to enable a conclusion, or at least indicate trends, may already exist.

In summary, in an alternating conditions design, conditions are unlikely to be systematically influenced by background variables, it can demonstrate control without being dependent on either stability or reversibility of behaviour, and thus it is a robust method for collecting data. Regardless of any pattern or inherent variability in behaviour, if subjects demonstrate consistently different responding under the alternate conditions, a conclusion about experimental control and relative effectiveness of the conditions can be made (Barlow et al., 1984; Ulman & Sulzer-Azaroff, 1975).

Interactions

In any design in which there are two or more conditions there is a possibility that these conditions will influence one another and confound the results (Watson, Singh, & Winton, 1985). These interactions have been labelled multiple-treatment interference (Campbell & Stanley, 1963), and condition-change interactions (Ulman & Sulzer-Azaroff, 1975), and may be troublesome on some occasions but useful and instructive on others (Hains & Baer, 1989).

Sequence Effects

One form of interaction among conditions is known as sequence confounding (Ulman & Sulzer-Azaroff, 1975), and occurs when the effect of previous conditions influences the effect of subsequent conditions. In a reversal design it is possible that any change (or lack of change) in behaviour during the experimental phase is not solely attributable to Condition B but is due, to some extent, to the initial influence of Condition A. For example, a client wishing to decrease the incidence of intrusive thoughts may be instructed to record each occurrence to establish a baseline. When treatment is implemented, however, its unique effects may be difficult to isolate if overt or covert recording behaviours were established during baseline and are not eliminated from the treatment phase.

An alternating conditions design has a greater likelihood of minimising sequence effects because conditions are presented only briefly (rather than for multiple sessions in one phase), so that any prolonged (interacting) influence is minimised, and each condition follows and precedes each other condition equally often, thereby counterbalancing any influences many times (Ulman & Sulzer-Azaroff, 1975).

More recently, it has been proposed that counterbalancing will not remove (and at best may only minimise) sequence effects, but that intelligent use of this technique within an alternating conditions design may make such influences more
obvious. A prototype design to accomplish this would alternate phases of each of
the conditions in isolation with phases in which the conditions are alternated.
Counterbalancing of conditions could be used both across and within phases. (A
full explanation of using an alternating conditions design to investigate interactions
can be found in Hains & Baer, 1989.) Although such a model is useful for
studying sequence effects, it is time and effort intensive, and these demands make
it unlikely to be useful in many realistic settings.

Carry-Over and Alternation Effects

There are two other types of interaction which may occur in alternating
conditions designs due to the rapid alternation of conditions: carry-over and
alternation effects. Carry-over effects occur when one condition influences another,
independent of their order of presentation (Barlow & Hayes, 1979; Watson et al.,
1985, discuss variations of this interaction), and can be considered as a sequence
effect which occurs within one phase (rather than between two phases) of the
alternating conditions design. Similarly, alternation effects describe interference
between conditions which is due to the rate of alternations and the length of the
interval separating exposure to each condition. Ultimately, both carry-over and
alternation effects are variations of sequence effects: one condition influences
another, and the rate of alternation is all that changes (Hains & Baer, 1989).

One way to deal with all sequence effects is to study the effect of each
condition both in isolation and in different paced alternations with the other
conditions of interest. This model, however, is not feasible in many laboratory and
applied circumstances because it is extremely demanding in terms of client and
researcher time and research resources. In treatment settings with humans, for
example, it probably should not be carried out because its lengthiness means that it
does not provide a prompt treatment intervention and it places methodological
investigation of interactions above client health (Hains & Baer, 1989).

The choice of single-subject design is ultimately dictated by pragmatic
concerns (Hains & Baer, 1989). When “ideal” designs are impractical it is
important to use techniques that should minimise carryover and alternation effects.
These techniques include providing clear discriminative stimuli to signal current
contingencies, random sequencing of conditions to distribute any carryover effect,
and employing sufficient delays between sessions to reduce or stop interference
between conditions (see Barlow & Hayes, 1979; Barlow et al., 1984; Barlow &
Hersen, 1984; McGonigle, Rojahn, Dixon, & Strain, 1987).
Design Considerations for the Present Experiments

In addition to the general advantages of the alternating conditions design already described, there were specific elements of the present experiments that supported the use of this design. First, neither the funds nor the time were available to complete "exhaustive" designs (discussed above) to investigate potential interaction effects. A second consideration concerns the unit of analysis. Because each frame within a set (of the programmed materials used in the current experiments) was considered to be of equivalent difficulty (similar step size), attempts were made during pilot work to gain experimental control by alternating conditions on each frame. Control was difficult to obtain in this way, however, and performance on each set was considered to be a more effective unit of analysis. Due to the finite length of the text being used (i.e., a total of 49 instructional sets), however, the maximum number of experimental comparisons possible was limited (e.g., a maximum of 24 comparisons of two conditions), and a design which offered good efficiency (number of experimental comparisons: number of sessions) was essential. Furthermore, an unavoidable consequence of using one set of the text as the unit of analysis is that, within each subject, performance can be expected to differ across sessions for reasons other than experimental manipulations. Pilot studies had shown that there was a substantial amount of "noise" in the data due to variability in content, set size, and difficulty of the materials used, so designs based on stability criteria were inappropriate. The Holland and Skinner text is undoubtedly of very good quality, but sets necessarily vary in the content of the material they present, their difficulty (which will depend on topic-specific entering knowledge), and their length (which may introduce fatigue as an additional uncontrolled factor). For example, in the first few sessions, subjects complete sets covering simple and conditioned reflexes (54 and 30 frames, respectively). Many sessions later they complete sets covering topics on stimulus discrimination (75 frames), chaining (37 frames), and the goals and techniques of science (33 frames). Not only do sets differ in the difficulty of the topic presented, but, as a result of previous training, different sets will vary in difficulty within one subject, and the same set will vary in difficulty across subjects.

Because of these unavoidable differences, the stable response levels typical of operant research in the laboratory are unlikely to be demonstrated with the academic behaviours being assessed in the following experiments. Anyone interested in studying academic behaviour or performance which varies due to changes in task difficulty, or anyone who must contend with unstable baselines, must consider the alternating conditions design (Ulman and Sulzer-Azaroff, 1975).
The alternating conditions design was used in the present experiments because it is likely to avoid problems with instability, nonreversibility, and premature termination, it is efficient in providing experimental comparisons, and it tolerates unstable behaviour levels. Interactions were possible but were minimised by using brief rapidly alternating sessions rather than prolonged ones (Ulman & Sulzer-Azaroff, 1975), by using clear discriminative stimuli (i.e., the experimental condition was labelled on screen), and by employing an obvious delay between conditions. Furthermore, variables that were sufficiently robust not to be diluted by interaction with others were of primary interest, and there was an intention to verify later any variables which artificially may be effective due to the presence of an interaction between conditions.

Computerised Programmed Instruction

Prior to the advent of the microcomputer, low cost, high power, readily accessible computers were not available. This hardware problem meant that most pre-1980 research on programmed instruction used textbooks as the presentation medium (Tudor & Bostow, 1991). There are several difficulties associated with using textbook presentations for experimental purposes: (a) students can peek at answers, or look back at previous materials (e.g., Anderson, Kulhavy, & Andre, 1971; Holland, 1960; Kulhavy, 1977); (b) strict control over the order and timing of presentation of stimulus materials, and presentation of reinforcement contingencies is impossible (Tudor & Bostow, 1991); and (c) measurement of dependent variables (e.g., response duration) is limited in precision (e.g., Kulhavy, Yekovich, & Dyer, 1976). The use of teaching machines in some studies provided better control over cheating and the presentation of contingencies, but, as mechanical devices, they still contained limitations for use as a research tool: They offered the experimenter limited control over their speed of operation, the presentation of stimulus material and contingencies, flexibility of manipulating stimulus parameters, and precision (particularly, in gaining detailed information on response characteristics).

Skinner (1986, 1989) maintained that the mechanical devices he used to implement his technology were primitive and that computers could teach more effectively. The computer is a much better tool because an appropriate controlling program written in a high-level language provides high-speed, consistent operation, there are no unprogrammed delays in delivering contingencies, there is the opportunity to modify experiment features with relative ease, and precise
measurements of response variables such as error counts and time spent on task are possible.

It is important to note that the following experiments represent a research effort that employs programmed instruction and computers as tools to examine the effect of external pacing. The experiments were not intended as an assessment of the efficacy of programmed instruction, nor were they designed to compare this strategy with other teaching approaches. Programmed instruction was used as a research vehicle because of its strong theoretical and empirical foundation, and no comparison between programmed instruction and other approaches was proposed.

Controlling Program

The controlling program for the present experiments was written in Turbo Pascal (version 5.5). To display characters on the screen the controlling program utilised assembly language routines (Crosbie, 1990) which are 1000 times faster than Pascal's standard WRITE procedure and present text on the screen within one video refresh cycle (i.e., in less than 10 ms). This was important because preventing delays in stimulus presentation is an important feature in experiments which require precise contingency structuring. A recent study (Tudor & Bostow, 1991) which presented programmed materials by computer, required 7 s loading time per frame, which is both lengthy, and introduces an apparently consistent, but unanalysed, variable into the task. Unprogrammed delays were clearly undesirable in the current experiments which aimed to assess the effects of programmed delays on learner behaviours.

Study Material

In 1958, Holland and Skinner used teaching machines to present an undergraduate course on behaviour analysis. The program consisted of 48 lessons, each made up of 29 frames, and was part of Skinner's first-year Natural Sciences course. Using records of student responses as feedback, the authors conducted detailed item-by-item analyses of the program, eliminated deficiencies, and revised and expanded the program. After several revisions, the program was longer, but took less time to complete, led to fewer response errors, and resulted in students making fewer self-scoring errors (Holland, 1960; Holland & Doran, 1973). The revised program consists of approximately 2000 frames and has been published as a programmed text: “The analysis of behavior: A program for self-instruction” (Holland & Skinner, 1961). Holland and Skinner (1961) was used to provide the study material for all experiments. The text is divided into 53 sets of between 25 to 75 frames. Of these sets, 49 are learning sets that present topics relevant to behaviour analysis, and the remaining 4 sets are review tests. The learning sets are
arranged into 14 conceptual parts. For example, Part 1 is entitled "reflex behavior" and this encompasses Set 1 (simple reflexes) to Set 6 (response mechanisms). Appendix 2-1 shows the table of contents from the programmed text (Holland & Skinner, 1961).

There were several reasons for using Holland and Skinner (1961): (a) the authors are experts in the subject matter and presentation mode, (b) the materials had been field tested with the target population, (c) the materials were developed and validated with experimental rigour and have a low blackout ratio (i.e., little irrelevant information; Kemp & Holland, 1966), (d) the text provided a large number of frames, and (e) the subject matter was not presented elsewhere in the undergraduate programs of the subjects used in this project. Although the published text contains 53 sets, the number used in the present experiments varied depending on methodological requirements.

A typist transcribed the materials from the text to the computer. Each set was compiled into a separate data file in which the frames were sequentially ordered (cf. the turn-the-page construction of a programmed text). The controlling program then accessed these files when necessary.

Exhibits

Many sets in the Holland and Skinner text begin with exhibits that provide reference materials for students as they work through those sets. Exhibits are either verbal (e.g., description of Pavlov's experimental design), diagrammatic (e.g., response patterns resulting from different reinforcement schedules), or both. For the present experiments these exhibits were photocopied, compiled in a presentation folder, and placed on the table with the computer. Instructions detailing which sets had exhibits to read, and when to read them, were supplied both in the folder and, at the appropriate times, by the computer.

Test Material

Experiments 3 to 6 employed a pretest, posttest, and follow-up test. The tests consisted of constructed-response items that were taken directly from the review sets included in Holland and Skinner (1961). In addition to the organisation described above, the text can be considered as consisting of four sections (Sets 1-16, 18-28, 30-40, 42-52), and Sets 17, 29, 41, and 53 review the preceding section. In addition to its association with the preceding section, each question in a review set is clearly labelled in terms of which learning set it assesses. Because this index was available, a pool of review items was established for which each item could be associated with a particular learning set and, hence, the experimental condition that was assigned to that set. Therefore, performance on any test item
could be related to the condition under which this material was learnt. This pool included only items that were relevant to the portion of the text that was presented to subjects. Eligible test items were arranged sequentially, and a computer program selected items (the actual number differed across studies) randomly without replacement.

**Role of Reviews in Programmed Instruction**

There are two review procedures in the implementation of the current materials that need to be clearly distinguished. Review sets were described in the previous two sections. The final set in each of the four sections of the Holland and Skinner (1961) text is a review set that assesses the material in the preceding section. The items in those review sets were used to make tests for the present experiments. Quite distinct from the review sets (that review the content of several sets using novel questions) is the review feature that is implemented with all other sets. This feature is described now.

One of the hallmarks of the original teaching machines was that a student mastered a topic before proceeding to the next. In Skinner's (1968) procedures, after the final question had been answered for a topic, the program returned to the questions that were answered incorrectly, and the subject reviewed those incorrect items until all questions were answered correctly. When commercial manufacturers started making teaching machines they decided that this feature was expensive to implement and subsequently abandoned it (Holland & Porter, 1961). Consequently, many modern CAI procedures do not have this feature. The review feature was incorporated in the experiments reported, however, because it has been found to improve performance by 10% (Holland & Porter, 1961), and the proposed experiments were performed to assess how programmed delays affect performance on an optimal implementation of programmed instruction.

In practice, the controlling program would record which frames received incorrect responses as students worked through a set for the first time. After the last item of a set had been completed, the program would then readminister the incorrectly answered items. For example, if a set had 29 frames and a subject responded incorrectly to frames 10, 12, and 25, these three items would be presented a second time, in numeric order, immediately after Frame 29 was completed. If any of the reviewed items were answered incorrectly, they would be presented in a second review, and this cycle would continue until all items had been answered correctly.
Self-Correction of Responses

Programmed instructional materials typically had students correct their own work. This is an unorthodox instructional procedure and was used because of the number and diversity of responses which students generate. Because many small steps are required to ensure successful directed learning, each student makes many overt responses. Overt (rather than covert) responses are critical because (a) they are easier to reinforce, and (b) the terminal behaviour being trained is the generation of discriminated verbal responses (as opposed to recognising and selecting one of several multiple-choice alternatives). When responses are constructed, however, there are often several equivalent answers to a question and it would be a waste of time (and particularly space, in textbook and machine versions of programmed instruction) to list them all (Holland & Skinner, 1961). Instead, the most commonly used correct answer is provided as feedback and students decide whether their answer is equivalent to this. Traditionally, students working in texts or on teaching machines created a paper record of all responses and corrections which were routinely checked for accuracy by a teaching assistant. Correction errors by students were infrequent (i.e., typically less than 2%; Holland, 1960).

Alternative response-correction procedures are available for constructed response items. Some CBI programs include software designed to recognise multiple acceptable responses and tolerate spelling errors and grammatical errors. This software is expensive, places high demands on computer memory, does not always detect correct answers, and decreases students' interaction with the materials (because they do not have to decide whether their response is sufficiently like the correct answer). In the present computer-based version of programmed instruction, the self-correction feature of programmed instruction was kept for several reasons: (a) modern software alternatives are expensive, memory intensive, and are not necessarily better; (b) self-correction necessitates attention to feedback which assists learning; and (c) it was important to maintain a close correspondence to the original technique and assess experimental variables in the best possible presentation of programmed instruction. In these experiments, subjects were instructed to use Holland and Skinner's guideline for self-correction: "use reasonable judgment in deciding whether your response is synonymous with the printed [screen-displayed] form. Score it correct if it is." (1961, p. viii).

Delay Duration

For the present studies, the duration of an imposed delay was 10 s. This duration was selected based on a consideration of the demands placed on the
learner by these materials, the findings presented in the literature, and pilot experience. It was suggested by one of the authors of the programmed materials employed in the present studies that lengthy delays would decrease performance due to distraction. Avoiding such delays is one of the advantages of a computer (cf. mechanical) implementation of programmed materials (J. G. Holland, personal communication, April 28, 1992). Hence, it was important to use the smallest duration delay that would still provide a learning advantage. The student typically read a couple of sentences and responded with one or two words. To reconsider this question, response and also the feedback, an imposed delay of 10 s seemed to satisfy the above criterion.

The results in a variety of delay studies have indicated that delays of a moderate duration typically lead to better performance than longer or shorter delays. For example, in Stokes et al.'s (1988) study of multiple-choice test performance, a 30-s delay group outperformed both a 0-s delay group, and a 60-s delay group. In Boehm et al.'s (1971) hospital location problem, delays of 5 min led to better performance than either delays of 8 min or no delays. Dyer et al. (1982) reported that response delays of 3 s to 5 s were more effective than 0 s or 6 s for improving discriminated performance of impulsive children. Clearly, the size of an effective delay will be task specific, but the principle of avoiding a delay that is too long or too short is relevant.

The programmed materials used in the present studies are characterised by small steps, discrete trials, and progressive building upon previous information. Similarities can be found Brackbill's two-choice discrimination task and Bourne's concept-identification task. Brackbill and her colleagues found that a 10-s feedback delay typically produced a DRE but that a shorter delay (5 s) could produce the effect when the material is difficult (see Brackbill et al.; 1967). Bourne and colleagues also showed that task complexity and available information determined the most effective duration for a postfeedback delay. On more difficult tasks, longer delays were better than shorter ones, but delays of between 10 s and 20 s were typically effective.

Finally, in pilot studies for the current project it was found that delays of 10 s were effective in improving learning, but that longer delays were no more effective. Hence 10 s was the delay duration chosen for use in the initial studies of this thesis.
Subjects

Subjects in the first experiment were students participating as part of their course requirements. All other experiments used paid volunteer students from Deakin University who responded to advertisements posted by the students’ association employment service. The advertisements stated that subjects were wanted for an experiment in human learning, and included the intended starting date of the experiment, its duration, the approximate length of sessions, and the amount of money that could be earned. Those students who contacted the laboratory for further information were given full details about the experiment, and an introductory session was organised in which the procedure was demonstrated and consent and employment forms were completed. Experimental sessions typically began the day following the demonstration. Prerequisites for participation were that subjects had English as a first language and had not previously studied the subject matter covered in the experiment. The number of subjects participating in each experiment and their demographic details are included in Appendix 2-2.

Apparatus

Laboratory

All experiments were conducted in the same 5.4-m x 4.4-m laboratory. The laboratory was windowless and generally very quiet. To provide an experimental space, which minimised distractions from other equipment in the laboratory, two 123-cm x 180-cm room dividers partitioned off a 180-cm x 200-cm area in one corner of the room. Subjects were seated at a 70-cm x 106-cm x 72-cm table with the computer and monitor to the rear of the table, and the keyboard and an exhibit folder directly in front of the subject.

Computers

Experiments 1 and 2 used a Samsung S330 (IBM PC compatible) computer with an ECM 36-cm EGA colour visual display unit. Experiments 3 to 6 used a Samsung S800 computer (IBM PC compatible) with a Samsung 36-cm VGA colour visual display unit. Both computers ran DOS version 3.3, a real-time, single-user operating system which is essential for presenting programmed delays.

Screen Display and Layout

A representative screen display is shown in Appendix 2-8. The background screen colour was black to provide good contrast, and most text was presented in primary colours to provide characters with clear boundaries and which do not pulsate (Isaacs, 1988). A single white line bordered the screen display. The set heading (e.g., Set 1: Simple Reflexes), question number (e.g., Question 1 of 54),
answer prompt ("Your Answer:") and score labels ("Correct", "Incorrect") were in green. Scores and the experimental condition label (e.g., "No delay between questions") were in yellow. Questions were presented in white, subject responses in red, and feedback (correct answer) in yellow. Instructions (e.g., "Please type your answer") were in cyan. In the original text version of the materials, important text (e.g., words, prefixes) was italicised. The computer presented this text in yellow. The screen could display characters in 25 rows (lines) and 80 columns. Questions were displayed in the centre of the screen (i.e., on Rows 12 to 18, and in Columns 20 to 60), although additional columns to the right were sometimes used to improve the appearance of a question. Students’ responses and the correct answer feedback were never more than one line of text, and appeared on Rows 20 and 21, respectively. Instructions were centred on Rows 23 and 24.

When a postfeedback delay was programmed, it was indicated on the screen by a long red bar ("Delay bar"). This was displayed centred on Row 24 (see Appendix 2-12), and indicated to subjects the proportion of the delay interval remaining, and that any typing during this time would not be processed. The Delay bar was 60 blocks (column widths) long and was overwritten in blocks of the background screen colour so that it gradually shortened at a uniform rate (10% per second), and such that it was entirely gone at the end of the delay period (10 s). The program would not accept any keyboard input if any part of the Delay bar was on the screen. Also, the keyboard buffer was cleared at the end of the delay to remove any characters typed during this time.

Tests

In Experiments 3 to 6, each subject completed a pretest, posttest, and follow-up test. The pretest was included to assess entering behaviour, and was presented on the day before subjects began the learning sets. On the day following completion of learning sets, subjects were given a posttest to provide a measure of retention. One month after the posttest, subjects completed a follow-up test to provide a measure of the extent to which the study material was maintained, and to indicate whether there had been any condition-specific decline in learning. Posttest and follow-up tests also resembled conventional classroom assessment techniques. Test construction is explained more fully in the relevant experimental sections.

Procedure

Introduction to the Task

When subjects arrived at the laboratory, they read a copy of the experimental instructions to gain familiarity with the task. The instructions were slightly different
in each experiment to accommodate changes in design and independent variables, so they are provided in full as Appendices 2-3 to 2-5. The instructions from Experiment 1 are representative:

**Experimental Instructions**

In this experiment you will use computer-assisted instruction to learn about an important area of Psychology: the experimental analysis of behaviour and its application. You will cover three topics per session, and there will be one session every day (Monday to Friday) for 15 sessions; each session will last between 1 and 1 1/2 hours.

Each topic has between 25 and 75 questions and for each you will supply one or two missing words. This study has three experimental conditions: (1) no delay between questions, (2) a 10-second delay after each question, and (3) a 10-second delay only after questions answered incorrectly. During a lesson only one of these conditions will apply (shown in yellow at the top of the screen). For the first five sessions you will receive only Condition 1 but for all other sessions you will receive all three conditions. You must answer all questions correctly before you finish a topic, so when you have answered the final question, the program will take you back to the questions you answered incorrectly, and will continue this process until all questions have been answered correctly. At the end of each lesson, you will be asked a question concerning the lesson, and for this you will type a number between 1 and 9 to show your answer.

Between sessions do not study the material or talk to other people about the topics; this might ruin the experiment, and make the results uninterpretable.

Subjects then worked through a demonstration with the experimenter and signed consent forms if they agreed to participate in the experiment. The demonstration was almost identical to a learning set except that it was greatly reduced in size and used a content entirely unrelated to the study materials. The demonstration material is described below. The consent form used in the first two experiments is included as Appendix 2-6. A modified consent form reflecting changes in experimental design and payment amounts, and requesting permission of use of academic results, was used in all subsequent experiments (see Appendix 2-7). For subjects in Experiments 1 and 2, the introductory session was then finished. For all other experiments, subjects also completed a pretest. Learning sets began the following day.
A Typical Learning Session

Each subject had a key disk that contained an executable version of the controlling program and a file showing the next lesson to be presented. To begin a session, students obtained their key disk (from a disk library next to the computer), inserted it in the computer, and typed ":begin." The program then read the condition file and loaded the materials for the current lesson from the hard disk into memory. This avoided the need for disk access (and the associated unprogrammed delays) during a session.

After a subject typed "begin", the screen was cleared before a message indicating the current lesson was presented. Subjects were prompted to check (in the exhibit folder) whether there was an exhibit for this session, and to press any key when they were ready to begin. These preliminary instructions read as follows:

This is lesson 1
If there is an exhibit for this lesson read it now.
Press any key when you are ready to start the lesson.

Once a key had been pressed, these preliminary instructions were cleared, and the first frame of the set was presented (illustrated in Appendix 2-8). When responding to the question presented in a frame, subjects were required to enter a response, self-score it as either correct or incorrect, and then press a key to move on to the next frame. Appendices 2-8 to 2-12 illustrate successive screens as a subject completed Frame 1 of Set 1 from Holland and Skinner (1961). After reading the question (Appendix 2-8), subjects typed their response (Appendix 2-9) followed by the ENTER key. Before pressing ENTER, the response could be altered using the BACKSPACE key. Once ENTER had been pressed, however, the response could not be changed, the correct answer was added to the screen (Appendix 2-10), and the instructions changed to:

Press C if your answer was correct
or I if it was incorrect

If any key other than "C" or "T" (or "c" or "t") was pressed the computer emitted a brief, low-pitched buzzing sound (300 ms, 100 Hz). Following a "C" or "T" (or "c" or "t") response, the score counter was updated, an incorrect frame was tagged by the controlling program for subsequent review, and one of the following two things would happen:
1. If it was a self-paced condition (i.e., no programmed delays) the following continuation prompt would appear (Appendix 2-11):

Press ESC. to stop, or any other key to continue

A further key press would result in presentation of the next frame.

2. If a postfeedback delay was programmed then it would appear at this point. The following instruction accompanied the Delay bar:

Please wait until the bar is gone

Appendix 2-12 shows the screen appearance when a Delay bar is present. It is important to note that the Delay bar was an addition to the screen so that during a standard implementation of programmed delays all other features (including the question, response, and correct answer) remained visible during this period. (However, in one condition of Experiment 2 all learning material was removed from the screen.)

After the Delay bar was completely overwritten by the background screen colour, the continuation prompt appeared (Appendix 2-11), and pressing any key initiated the next frame. After the final frame in a set had been attempted by a subject, those frames for which responses had been incorrect were repeated (review feature). The screen appearance and sequence was not different for repeated items. In Experiments 1 and 2, repeated items were presented under the same conditions as they were during their first presentation (e.g., accompanied by a programmed delay). In subsequent experiments, however, repeated items were all presented under no-delay (baseline) conditions.

Another noteworthy aspect of the present procedures was that in both no-delay and delay conditions the subject always initiated each stage of the frame (i.e., making responses, correcting responses, and initiating the next item), and external control of pacing occurred only once per frame during delay conditions, when the 10-s delay was in progress.

After all frames had been completed correctly the screen was cleared and a question designed to gauge whether subjects liked the experimental condition appeared centered on the screen (displayed in yellow on Row 8); this was accompanied by a condition reminder (cyan on Row 12) and response key (red on Row 14):
How satisfied were you with the previous experimental condition?

No delay between questions

1 = not at all satisfied . . . 9 = extremely satisfied

The condition reminder was always the condition label of the most recent set. After a subject typed a number between 1 and 9 (which was echoed in the centre of Row 16) and pressed ENTER, the controlling program cleared the screen, and then did one of two things: (a) If there was another set to be completed in this session (as may have occurred in Experiments 1 and 2), the subject was again presented with the preliminary instructions (i.e., “This is lesson 2”, etc.); or (b) wrote the word “Finished” in cyan in the centre of the screen, to indicate that the session was complete.

At the end of an experimental session, the controlling program incremented the next-session number in the condition file, and wrote relevant session details to an output file. These details included, for each frame, the subject’s answer, the feedback, and the subject’s correction (correct or incorrect). In Experiments 3 to 6 timing measures also were recorded for each frame. In addition to frame details, summary statistics were written to the end of the file, and these included the total number of frames in the set, the number of repeated items, the number (and percentage) of incorrect items, the timing variables measured for that experiment, and the satisfaction rating.

An experimenter was always available in the laboratory to deal with any difficulties encountered by subjects or problems with the equipment.

Demonstration

Prior to the beginning of an experiment, subjects interacted with a 9-frame demonstration that introduced the general nature of the task, important features of the screen display, and notation conventions used in frames. Screen displays and sequence used in the demonstrations were identical to those used in the experiments, but the content of the frames was unrelated to the study materials. All relevant features of the screen display (see, for example, Appendix 2-8) and sequence were demonstrated: title of the present topic, number of frames to be completed, current experimental condition, score box, constructed responses, reviews and, if delays were present in the forthcoming experiment, the Delay bar (see Appendix 2-12). The demonstration also was used to explain several frame conventions: (a) “_____” required a one-word response, (b) “_____  _____”
required two words, (c) "***" required as many words as was necessary to respond to the item, (d) "_____ (TT)" indicates that a technical term was required, and (e) highlighted text (in yellow instead of white) represented important information. The demonstration also gave subjects who were computer naive an opportunity to interact with the keyboard and display features, and to appreciate that no background in computing was necessary to participate.

**Subject Self-Scoring**

In the present experiments, each subject's responses and their corrections were recorded by the controlling program. For each session, these records were analysed by the author to assess the extent to which subjects adhered to the self-scoring guideline. The number of frames that were marked as being correct when they were incorrect, or marked incorrect when they were correct (this occasionally happened), was summed, divided by the total number of frames completed (corrected) for that set, and then converted to a percentage. The percentage of self-correction errors for each subject is reported in Appendix 2-13.

**Experimental Timetable**

In Experiments 1 and 2, subjects attended the laboratory for one session per day for 15 consecutive weekdays, and debriefing immediately followed the final session. The series of events for Experiments 3 to 6 was as follows: On Day 1, subjects completed the pretest; on Days 2 to 21 they completed the learning sets (2 sessions per day for 20 consecutive weekdays, and these sessions were separated by a minimum of 2 hrs); on Day 22 they completed the posttest; and, approximately one month later, subjects completed the follow-up test. Debriefing occurred immediately after the follow-up test.

For all experiments, weekly timetables were posted on the laboratory door to enable subjects to organise their own session times around their studies and avoid clashes with other subjects.

**Tests**

Tests were initiated by subjects inserting their disks and typing "test". Test items were then presented by a version of the controlling program which was modified in the following way: It did not present the score box on the screen display, or provide feedback to the subject; furthermore, self-correction procedures and social validation questions were eliminated. After a response had been entered and ENTER pressed, the controlling program immediately presented the next frame. There were no reviews or delays in tests. When all questions had been completed, the controlling program incremented the number in the condition file which represented the subject's next test, and then wrote the relevant session
details to an output file. These details included the subject’s answer, the correct answer, and timing measures.

**Debriefing**

Following the completion of each experiment, subjects were debriefed. Debriefing involved informing subjects of the historical role of automated instruction, and the practical relevance of a better understanding of the processes involved in instruction and learning. They were told the aims of the current experiment as well as the general purpose of the different experimental conditions, and were asked not to discuss the aims and experimental techniques with other students. Subjects also were asked questions about experimental events (e.g., What was the purpose of the different conditions? Which condition was preferable, for what reasons? Which condition was more difficult to work in?). Subjects’ general comments can provide important information about the experiment such as problems with the controlling program (e.g., faulty disks causing unprogrammed delays), adventitious contingencies, and unforeseen confounds.

After debriefing, subjects in Experiments 3 to 6 completed two evaluative questionnaires (Moore & Smith, 1964; Van Atta, 1961) which obtained general feedback about automated instruction.

**Analyses**

**Dependent Variables**

Throughout the experiments a variety of behaviour measures were used. There were, however, three central behaviours of interest: accuracy, efficiency, and satisfaction. Accuracy was operationalised as the percentage of frames answered correctly in the first pass through a set. Efficiency refers to a subject’s mean time per frame in any given session and was calculated by dividing the total time (including programmed delays and repeats) required to complete a set by the number of frames in the set. Hence, this measure reflects the inefficiencies of repeated items and external delays. Satisfaction was determined by having subjects use a Likert scale (1 = not at all satisfied . . . 9 = extremely satisfied) to rate, at the end of each set, their satisfaction with the experimental conditions experienced during that set.

Although specific hypotheses are reserved for each experiment, some general expectations about the influence of programmed delays on these performance measures can be outlined here. Accuracy was expected to increase in the presence of delays. Efficiency was expected to decrease due to delays, but only by a small amount: Increased accuracy (hence, fewer repeats) and quicker discriminated
responses should offset some of the time lost due to imposed delays. It was expected that subjects would rate satisfaction with delay and no-delay conditions similarly. However, because session-by-session ratings of satisfaction may have been influenced by a variety of secondary factors such as performance, fatigue, and lesson content, global opinions of the conditions were always canvassed during debriefing to provide supplementary information.

**Visual Inference**

Visual inference is the predominant method of evaluating experimental manipulations in single-subject designs (Kazdin, 1982; Sidman, 1960), and for this reason was employed in the following experiments. Traditionally, single-subject data are obtained from animal subjects in laboratory experiments. In this paradigm, tight control over experimental conditions is possible and, if necessary, many sessions could be employed to obtain stable rates of responding. The influence of an experimental variable is judged by visual inspection of response rates before and after its introduction. It is considered that a significant effect occurred if response rates changed to a level clearly different from baseline levels, if this change occurred immediately after the introduction of the new conditions, and if this effect remained during the presence of the experimental variable. Without clear visual evidence, it could be concluded that either the variable did not substantially influence the behaviour or that sufficient control had not been obtained. These criteria are also useful in some human experimentation; in an applied or clinical domain, intervention effects typically must be large and dramatic to be of practical value, and such potent effects will also be obvious by visual inspection.

Visual inference of an experimental effect can be problematic, however, particularly in applied research with human subjects. There is often limited control over subjects’ history and background variables which contribute to variability in behaviour levels. Furthermore, short baselines are often used (Sharpley, 1987) which can make stability and trend difficult to interpret, and undermine confidence in judgements of change in the experimental phase. Clear, substantial effects are not ubiquitous and research conducted with more typical operant data has found that visual inference is both unreliable (DeProspero & Cohen, 1979; Furlong & Wampold, 1982; Jones, Weinrott, & Vaught, 1978) and unable to control Type I error (Matyas & Greenwood, 1990), and it is considered a relatively insensitive (i.e., only detects marked effects) method for determining if a treatment has a significant effect (Barlow & Hersen, 1984). It must be noted that this evidence has been obtained from research evaluating judgements of relatively simple AB and
ABAB designs (A = baseline, B = intervention), and may not apply to findings from an alternating conditions design as used in the present thesis. Visual inference in the latter case may not be problematic, but, for the purposes of the present experiments, this cannot be assumed.

Statistical Analyses

In light of the problems associated with visual inference, many researchers have advocated the use of statistical procedures to aid the decision process (e.g., Matyas & Greenwood, 1990). It has been proposed (Kazdin, 1976) that in situations in which (a) stable baseline data cannot be achieved; (b) when “the data may appear to reveal reliable effects, although visual inspection is equivocal” (p. 270), such as assessment of variables whose effects are not well known; or (c) when there is variability due to extraneous factors which cannot be controlled, statistical techniques can be used to assess reliable effects. One or more of these situations was likely in the current paradigm in which there were four known sources of variation: (a) set size, (b) set difficulty, (c) set content, and (d) the experimental variable. Set size, difficulty, and content were uncontrollable factors which were likely to introduce considerable variability into the data, and make stability unlikely. If this occurred it would lessen the visual impact of the experimental variable which was expected to have a consistent effect, but one that may have been small relative to background “noise.” Hence, to support inferences drawn from visual inspection, some statistical analyses were used.

The Present Data

During the alternating conditions phase in each experiment, each subject provided a series of matched data pairs: Hypothetical data are presented in Figure 2-2.

<table>
<thead>
<tr>
<th>Day</th>
<th>Sets</th>
<th>Delay</th>
<th>No Delay</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>31 &amp; 32</td>
<td>78.0</td>
<td>72.3</td>
<td>5.7</td>
</tr>
<tr>
<td>12</td>
<td>33 &amp; 34</td>
<td>88.0</td>
<td>80.5</td>
<td>7.5</td>
</tr>
<tr>
<td>13</td>
<td>35 &amp; 36</td>
<td>61.3</td>
<td>89.9</td>
<td>-28.6</td>
</tr>
<tr>
<td>14</td>
<td>37 &amp; 38</td>
<td>87.5</td>
<td>79.8</td>
<td>7.7</td>
</tr>
</tbody>
</table>

Figure 2-2. Hypothetical data provided by one subject across four days, using an alternating conditions design presenting two conditions (No Delay and Delay) per day.

On Days 11, 12, and 14, the subject scored more studying under Delay than No Delay. Unfortunately, large differences in the opposite direction to a relatively
consistent pattern (such as occurs in Set 36 on Day 13) may result from other variables such as size, content, or difficulty of set. This is an inherent difficulty that arises from working with materials that are educationally realistic. If the experimental variable can demonstrate a consistent effect above and beyond such variability then there is good reason to be confident in its controlling ability. Because such “spurious” data pairs can arise, however, it is important to employ analyses which are sensitive to consistency of effect, rather than an average magnitude across sessions.

Parametric tests are inappropriate for these data because such tests assume a normal sampling distribution. This requires that either the population scores are normally distributed, which is unlikely, or that the sample size is large, which is not the case. Furthermore, tests which compare mean differences between conditions (e.g., repeated measures t-test) emphasise the overall magnitude of differences across conditions. Overall condition differences can be greatly influenced by large variability in set scores, and this masks small consistent differences.

Because nonparametric tests make few assumptions about population parameters (Kaplan, 1987) they were considered more appropriate for the current data. Several nonparametric tests are available for single-subject multiple-condition designs (Edgington, 1982), but the sign test (Siegel, 1956) was considered the most appropriate of these because it is based on the direction of the difference in measurements rather than the magnitude, it has the fewest assumptions, and it is conservative. For each matched data pair, the sign test considers the direction of the experimental effect: if the effect occurs in the same direction with high frequency the test considers the treatment to have had a significant effect. The magnitude of the difference between values in a matched pair are not taken into account, nor is the variability in values across pairs. Applied to the present experiments, if the experimental variable (e.g., noncontingent postfeedback delay) produces changes in behaviour consistently enough to override the influence of set variability, it is clear evidence of behavioural control exerted by this variable.
CHAPTER 3: EFFECTS OF ENFORCED DELAYS

Experiment 1

The aim of this experiment was to assess the relative effectiveness, efficiency, and social validity of self pacing and external pacing in programmed instruction. Self pacing was operationalised by providing students with complete control over their rate of progress through the materials (No Delay). External pacing was operationalised in two ways. In the first way, brief, noncontingent delays were imposed following feedback on each trial (Noncontingent Delay). Postfeedback delays were employed because it is at this point that learners have the maximum amount of relevant information available. In the second form of external pacing, brief, response-contingent delays (presented only after incorrect responses) were imposed between feedback and the following stimulus (Contingent Delay).

Although noncontingent delays were the primary focus in this experiment, there were important reasons to assess contingent delays. In addition to providing a control for noncontingent delays, contingent delays may function as a form of punishment. Punishment for incorrect responses is an extremely effective way to produce discriminated performance (Gestie, Langer, & Glass, 1985). This has been shown consistently with laboratory animals (Fowler, Hochhauser, & Wischner, 1981), intellectually disabled children (Harris & Tramontana, 1973), intellectually normal children (Miller, Moffat, Cotter, & Ochocki, 1973), impulsive children (Hemry, 1973), hyperactive children (Cunningham & Knights, 1978), and adults (Matthews & Shimoff, 1974). When punishment is employed in discrimination training, subjects respond more slowly, and attend more carefully to stimulus features (Barlow, 1933; Muenzinger, 1934a, 1934b). It is, therefore, reasonable to expect that punishment may also improve performance with programmed instruction.

Skinner’s principles of programmed instruction and his two basic tenets that students should select their own pace, and study contingencies should be positive, have been described in Chapter 1 (in section Automated Instruction; Skinner). Noncontingent Delay provides a direct assessment of the first of these tenets. Contingent Delay serves a dual function of being another form of external control and, if it operates as a punisher, will allow an assessment of the second tenet of programmed instruction.
Method

Subjects

Four female undergraduate students served as subjects: H21 was a 30-year-old humanities student, H22 was a 23-year-old psychology student, H23 was a 23-year-old social sciences student, and H24 was a 20-year-old humanities student (see Appendix 2-2 for subject details). All subjects had completed one course in applied behaviour analysis (based on Miller, 1980), but had not seen the material used in the present experiment. All subjects were naive with respect to the aims and rationale of the present experiment and participated as part of the course requirements of a third-year research project. Their grades for this course were not contingent on their performance, but, because their grades were contingent on their writeup of the experiment, it was expected that they would be highly motivated.

Apparatus

Design. Each subject received the following conditions: (a) no delay between items (No Delay), (b) a 10-s delay after each item (Noncontingent Delay), and (c) a 10-s delay after items answered incorrectly (Contingent Delay).

Three sets were completed each session, and each set took approximately 20 min, so the 45 sets were completed in fifteen 1-hr sessions. For the first five sessions, the three sets used in each session were presented with No Delay to establish a baseline against which to compare the effects of the two delay conditions, and to adapt subjects to the equipment and procedures before other conditions were introduced. Following baseline sessions, there were 10 sessions and each included three sets: one set with No Delay, one set with Noncontingent Delay, and one set with Contingent Delay. The assignment of conditions to sets and the presentation sequence of conditions was determined by a random sequence with the constraint that an identical sequence of conditions could not occur in consecutive sessions. The assignment of conditions to sets is presented in Appendix 3-1. This arrangement resulted in 1112 frames being used in the alternating conditions phase with No Delay, Noncontingent Delay, and Contingent Delay consisting of 358, 404, and 350 frames, respectively. In the present experiment each subject received the same sequence of conditions, so their results were directly comparable. This also was true within all subsequent experiments. Across experiments, however, the sequence of conditions sometimes changed.

Procedure

On the day prior to the first learning session, subjects read the experimental instructions, interacted with the demonstration program, and completed consent forms. On Days 1-5 subjects completed the baseline sets; on Days 6-15 they
completed the alternating conditions phase. After the final session they were debriefed. Refer to the General Method for further details of these procedures.

When working through a set, for each of the questions presented on the computer screen, subjects provided constructed responses (i.e., fill-in-the-blank response mode), read the correct-answer feedback, and then pressed either “C” or “T” to record their response as correct or incorrect, respectively. During No Delay, immediately after a response had been scored a subject was instructed to type any key to continue. The current question, response, and feedback remained visible until a key was pressed, then the screen was cleared and the next frame was presented. During Noncontingent Delay, immediately following a subject’s correction, and regardless of whether the response was correct or incorrect, the Delay bar appeared at the bottom of the screen to indicate that the subject had lost control of pacing until the Delay bar was gone. For Contingent Delay, delays were programmed to occur only after responses scored as incorrect. During delay trials (i.e., for noncontingent delays and contingent delays) the Delay bar was an addition to the screen, so all other features (including the question, response, and correct answer) remained visible during the delay period. Throughout the delay period the keyboard was locked so that typing had no effect and, immediately after the delay, the keyboard buffer was cleared of any characters typed during this period. After a subject had attempted all questions in a set, those answered incorrectly were presented again as part of a review feature. Reviews continued until every item had been answered correctly once, and the experimental contingencies were maintained throughout the reviews (i.e., delay trials that were incorrect were repeated under delay conditions).

In each condition the subject always initiated each stage of a frame (i.e., making responses, correcting responses, and initiating the next item), and external control of pacing occurred only once per frame during the delay conditions, when the 10-s delay was in progress.

**Results**

Because sets differed in terms of number of frames, content, and difficulty, the number of correct responses per set and the time taken to complete a set sometimes varied markedly from session to session. Hence, results were standardised to permit a fair comparison across conditions; the following analyses are based on the percentage of items answered correctly during the subjects’ first pass (i.e., excluding reviews) through each set, and the mean time per item (including time due to programmed delays and repeated items). Satisfaction ratings also were compared across conditions.
The data for each dependent variable are presented in three forms:

1. A figure displaying the session-by-session data. This is the traditional way to present behavioural data, and such figures can display any consistent differences between conditions relative to the normal variability in behaviour level. The present data, however, often show variability in performance (reflecting the variability in the materials) that can obscure differences between conditions. To assist comparisons, each graph shows only two conditions: Noncontingent Delay versus No Delay (left column), Noncontingent Delay versus Contingent Delay (centre column), and No Delay versus Contingent Delay (right column). The figures show subjects' performance for the last five of the fifteen baseline sets to save space;

2. A table of binomial comparisons that shows any consistent differences in performance between conditions during the alternating conditions phase. These statistical comparisons are used to provide an index of the consistency of effects, and to support the visual presentations, but are not a major emphasis of this analysis, so no attempt was made to control Type I error when several comparisons were made; and

3. A figure showing each subject's performance in each condition averaged over sessions. This figure summarises the raw data, reduces daily and individual variability, shows the magnitude of any experimental effect over the entire experimental phase, and also shows the consistency of this effect across subjects.

**Accuracy (percentage of frames correct)**

Figure 3-1 shows the percentage of frames answered correctly for the last five sets of baseline and each set of the alternating conditions phase. The first two columns suggest that accuracy is typically greater under Noncontingent Delay (filled circles) than No Delay (unfilled circles), and Contingent Delay (filled triangles), respectively. The third column suggests no consistent advantage for either No Delay or Contingent Delay. Hence, this figure suggests that Noncontingent Delay leads to greater accuracy than either of the other two conditions.
Figure 3-1. For each subject, the percentage of frames correct per session (excluding reviewed items), for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Contingent Delay (CD; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison.
Table 3-1
Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Percentage of Frames Correct than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; CD</th>
<th>ND &gt; CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>H21</td>
<td>9/10*</td>
<td>9/10*</td>
<td>6/10</td>
</tr>
<tr>
<td>H22</td>
<td>5/10</td>
<td>6/10</td>
<td>5/10</td>
</tr>
<tr>
<td>H23</td>
<td>8/10</td>
<td>7/10</td>
<td>6/10</td>
</tr>
<tr>
<td>H24</td>
<td>7/10</td>
<td>6/10</td>
<td>4/10</td>
</tr>
<tr>
<td>Total</td>
<td>29/40**</td>
<td>28/40**</td>
<td>21/40</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and CD = Contingent Delay.
* p < .05. ** p < .01.

Figure 3-2. Percentage of first pass frames answered correctly in baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Contingent Delay, CD) for each subject. Median results across subjects are also shown.
The figure also shows that accuracy was reasonably stable during baseline and increased in variability during the alternating conditions phase. This increased variability may have been due to an interaction among the conditions (as sometimes occurs with multielement designs: Hains & Baer, 1989; Sidman, 1960), or differences in set difficulty. The extent to which an interaction existed cannot be determined. Consideration of the materials suggests, however, that set difficulty did become more variable. The text is conceptually divided into fourteen parts (e.g., Part 1 Reflex behavior), and each part consists of two or more sets (see Appendix 2.1).

When a new conceptual part is introduced, earlier sets teach more elementary information, whereas more difficult material is covered in later sets. In addition to this increasing complexity within each part, the text tends to cover more complex issues as it progresses, and later concepts depend on earlier concepts by differing amounts.

Table 3-1 shows for each subject, and scores aggregated across all subjects, the consistency across sessions of differences in accuracy between conditions during the alternating conditions phase. In a sense, it is the numerical summary of the consistency of the effects shown in Figure 3-1.

Because the conditions were comparable except for the experimental contingencies (Sidman, 1960; Ulman & Sulzer-Azaroff, 1975), a sign test (Siegel, 1956) was employed to determine the probability that one condition would have a consistently higher score than another condition by chance (see Edgington, 1982, for a justification of this statistical procedure). For example, across all subjects, Noncontingent Delay had a higher percentage of frames correct than No Delay on 29 of the 40 comparisons on which their scores differed. Assuming an initial probability of .5 (i.e., each condition is likely to have a higher score than the other on 50% of comparisons), the probability that such an extreme result would occur by chance is less than .005 (cumulative sign test). This pattern also was evident for three of the four subjects (H21, H23, and H24). Results for individual subjects are presented primarily to show consistency of results across subjects.

Across all subjects, the percentage of frames answered correctly for Noncontingent Delay was greater than Contingent Delay for 28 of the 40 sessions, and this difference is also significant (p < .01, sign test). There was no systematic difference in accuracy between No Delay and Contingent Delay.

Figure 3-2 shows, for each subject and condition, the percentage of frames answered correctly during the alternating conditions phase. For all subjects, Noncontingent Delay produced more correct responses than did the other
conditions. In addition to the data from each of the four subjects, the median results across subjects are shown. Median results show that Noncontingent Delay produced approximately 10% more correct responses than No Delay, and approximately 6% more correct responses than Contingent Delay. There was some suggestion of a ceiling on the potential difference between No Delay and Noncontingent Delay: When accuracy in No Delay was high, Noncontingent Delay led to only small improvements (see H22), but when accuracy in No Delay was lower, Noncontingent Delay produced greater improvement (see H21, H23). There also was evidence of a relationship between the amount of delay received and the extent of improvement in accuracy: Median data show that accuracy was best in the condition in which delays occurred most frequently (Noncontingent Delay), worst when delays were not employed (No Delay), and intermediate when delay was presented intermittently (Contingent Delay). This pattern where accuracy corresponds to amount of delay received was present in three of the four subjects (H21, H22, and H24).

Reliability of self scoring. Because subjects scored their own responses (i.e., indicated whether a response was correct or incorrect), their reported accuracy could be inaccurate, or worse, biased in favour of particular conditions. The record was checked for each subject and session to assess this possibility, and it was found that all subjects had an accuracy rate greater than 98% (see Appendix 2-13), and there was no difference in accuracy across conditions. In addition, when the analysis was repeated with the corrected data, the same pattern of results occurred. Hence, the reported scores can be used with confidence. The present accuracy figure also adds generality to previous findings (based on the same materials) that students usually score over 98% of responses accurately (Holland, 1960).

Review sets. Three of the 30 sets employed in the alternating conditions phase were reviews of previous sets. Hence, it was possible that results for these review sets were affected not only by the experimental condition operating during the review, but also by the condition in operation when each question was initially presented in a prior set. In an attempt to overcome this possible confound, the comparisons shown in Table 3-1 also were made only for the seven sessions which did not contain review sets. These comparisons produced the same pattern of significant results as those shown in Table 3-1. Hence, the use of review sets did not affect the present results, and ensured that the present procedures were directly comparable with the textbook version of Holland and Skinner (1961).
Efficiency (mean time per frame)

Mean time per frame was calculated by dividing total time (including delays and repeats) required to complete a set by the number of frames in the set. Hence, this measure reflects the inefficiencies of repeated items and external delays.

Figure 3-3 shows that mean time per frame in each session was reasonably stable during baseline (except for Session 3 for H22), but increased in variability during the alternating conditions phase. In none of the three columns does there appear to be a consistent difference between the two conditions presented. Table 3-2 also shows that there were no consistent differences between any of the three conditions in the efficiency with which subjects completed sets.

Figure 3-4, on the other hand, shows that Noncontingent Delay had a longer mean time per frame than the other conditions for all subjects. Medians show that frames completed under Noncontingent Delay took 4 s per frame longer than No Delay, and 6 s longer than Contingent Delay. These data show that there is an overall time cost associated with the increased accuracy under Noncontingent Delay. They also show, however, that subjects must have saved time in some way under Noncontingent Delay because the average time cost per frame was less than the 10 s added to each frame.

At first sight, the findings presented in Table 3-2 and Figure 3-4 may appear contradictory: the former implies that there are no differences between conditions in mean time per frame, whereas the latter suggests that Noncontingent Delay is less efficient than the other two conditions. The table and figure, however, present different aspects of subjects' behaviour. The table of binomial comparisons shows any consistent advantage of one condition over another, whereas the bar graph shows the magnitude of any overall differences between conditions, regardless of any consistent effect. Together, the presentations show that sessions completed under Noncontingent Delay typically did not require more time, but in some sessions they did, and this contributed to longer times on average.
Figure 3-3. For each subject, mean time per frame per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Contingent Delay (CD; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison.
Table 3-2
Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Mean Time per Frame than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; CD</th>
<th>ND &gt; CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>H21</td>
<td>5/10</td>
<td>6/10</td>
<td>5/10</td>
</tr>
<tr>
<td>H22</td>
<td>7/10</td>
<td>7/10</td>
<td>6/10</td>
</tr>
<tr>
<td>H23</td>
<td>6/10</td>
<td>6/9</td>
<td>5/10</td>
</tr>
<tr>
<td>H24</td>
<td>6/10</td>
<td>6/10</td>
<td>5/10</td>
</tr>
<tr>
<td>Total</td>
<td>24/40</td>
<td>25/39</td>
<td>21/40</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and CD = Contingent Delay.

Figure 3-4. Mean time per frame during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Contingent Delay, CD) for each subject. Median results across subjects are also shown. Weighted means have been used because varying set sizes and random assignment of conditions to sets resulted in slightly different numbers of frames in each experimental condition.
Satisfaction

Figure 3-5 shows that there was a moderate to high level of satisfaction during baseline, but it was highly variable. In the alternating conditions phase, variability continued and there was no obvious change in level of satisfaction. Although one subject (H22) showed a clear preference for No Delay, then Contingent Delay, then Noncontingent Delay, there were no consistent preferences for any of the conditions in the remaining three subjects (H21, H23, and H24).

Table 3-3 shows that, across all subjects, there were no differences in ratings of satisfaction between Noncontingent Delay and either of the other conditions. There was an overall preference for No Delay over Contingent Delay (p < .05, sign test), and three of the four subjects showed this preference. Comparisons for H22 confirm the preferences which were visible in the raw data of this subject.

Figure 3-6 presents the mean satisfaction rating for each subject in each condition plus median results across subjects. Median results show that all conditions had a moderate satisfaction rating (approximately 5 on a 9-point scale), and that No Delay was rated slightly higher (0.1) than Noncontingent Delay which was rated 0.2 higher than Contingent Delay. Delays did not distress subjects. The most consistent result was that three of the four subjects (H21, H23, and H24) preferred Noncontingent Delay to Contingent Delay.

Additional Findings

Absolute number of errors. Figure 3-2 shows that the percentage of items answered incorrectly was approximately 10% for H22, 30% for H21 and H24, and 40% for H23. Because the expected error rate for the present materials (and most programmed materials) is regarded as 10% (Holland, 1960; Holland & Doran, 1973; Holland & Skinner, 1961), the current figures may appear to be too high. Several factors, however, must be considered when making this assessment:

1. When the present materials were developed, there were circumstances which probably contributed to excellent performance (J. G. Holland, personal communication, April 28, 1992): (a) students from Harvard and Radcliffe served as subjects in those studies and they were outstanding students; (b) they were examined on the materials so there was a strong contingent relationship between study behaviour (accuracy in learning sets) and grades, whereas all students in the present experiment received identical participation credit that was not contingent on accuracy; and (c) there was likely to have been considerable competition between those students.
Figure 3.5. For each subject, satisfaction per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Contingent Delay (CD; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison.
Table 3-3
Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Satisfaction than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; CD</th>
<th>ND &gt; CD</th>
</tr>
</thead>
<tbody>
<tr>
<td>H21</td>
<td>7/9</td>
<td>5/8</td>
<td>3/7</td>
</tr>
<tr>
<td>H22</td>
<td>1/10*</td>
<td>1/9*</td>
<td>6/6*</td>
</tr>
<tr>
<td>H23</td>
<td>3/6</td>
<td>6/9</td>
<td>6/9</td>
</tr>
<tr>
<td>H24</td>
<td>3/6</td>
<td>5/8</td>
<td>6/9</td>
</tr>
<tr>
<td>Total</td>
<td>14/31</td>
<td>17/34</td>
<td>21/31*</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and CD = Contingent Delay.
* p < .05.

Figure 3-6. Mean satisfaction per session during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Contingent Delay, CD) for each subject. Median results across subjects are also shown.
2. It is important to realise that the accuracy guideline of 10% error is a traditionally used target error rate for a frame, not an entire set; a program author will modify a frame if more than 10% of subjects answer it incorrectly. The present data refer to subject's average accuracy across sets which is not unlike that found in typical undergraduate courses: it represented grades from A (H22) through to D (H23).

3. In a recent study of computerised programmed instruction published in *Journal of Applied Behavior Analysis* (Tudor & Bostow, 1991), subjects in the condition most similar to the present procedures (Group 5) had an average error rate of over 30%. For the above reasons, the present error rates do not appear to be excessive.

**Set completion times.** Set completion times of the present subjects were greater than estimated completion times published in the original text. Again, this difference appears to be due to the standard of the students on which those estimates were based, and time pressures associated with completing materials before examination. In fact, the estimated times required to complete a set provided in Holland and Skinner (1961) actually represent a great range of times (cf. Holland, 1960) and a quicker pace than is commonly attained by students (J. G. Holland, personal communication, April 28, 1992).

**Discussion**

The present experiment assessed whether noncontingent and response-contingent postfeedback delays affect performance on programmed instruction. Response-contingent delays had no consistent effect, but noncontingent delays consistently increased accuracy by approximately 10%, with no decrease in subjects' satisfaction, and only a small overall decrease in efficiency.

Contingent Delay did not differ from No Delay in either efficiency or satisfaction, but did result in a small, but inconsistent, increase in percentage of frames answered correctly. This suggests two possibilities: (a) that a 10-s delay following an incorrect response is not different from No Delay; or (b) because it was presented intermittently, its influence did not occur sufficiently often to be detectable by the statistical test employed in this experiment. Because there is no evidence that Contingent Delay operated as punishment, it is imprudent to infer anything concerning the effects of punishment in computerised programmed instruction. Future studies could assess this issue further by employing a larger range of response-contingent delays or response-cost contingencies.

There are other instances of unsuccessful use of contingent-delay procedures in programmed instruction. In one study (Anderson et al., 1971), 15-s
response-contingent delay periods were employed as a form of mild punishment designed to increase carefulness. Compared to a condition closely resembling No Delay, 15-s delays did not improve accuracy, retention, or first pass time. Even if response-contingent delays can be increased to a duration that consistently improves accuracy, the time cost involved may be prohibitive.

Although little can be concluded concerning Contingent Delay, the present experiment did provide an unambiguous assessment of the effects of providing noncontingent delays after each item. Compared to a condition in which there was no external control of pacing, subjects experiencing consistent, brief delays completed programmed materials more correctly, with no loss in satisfaction, and only a small increase in time taken. Hence, the findings from the present experiment are contrary to the doctrine that student control of pacing optimises learning. It is also ironic that speed, one of the main putative advantages of computers compared with mechanical teaching machines (Skinner, 1986), can impair accuracy.

Why did external pacing improve accuracy? Two possible explanations of the improved accuracy in Noncontingent Delay are:

1. Extra-study proposition: Delay periods allow extra study time. In the present experimental design, the item, correct answer, and subject’s response were all visible on the screen during the delay period. It is possible that subjects inspected the material during this time to consolidate, or establish, correct discriminations. Increased accuracy may be a result of having spent more time on the materials (Tobias, 1973).

2. Racing proposition: In some sessions, students mentioned that they hurried because they had other commitments outside the laboratory. We have also observed “hurrying” in other classes where students work on computers to complete course materials, and it appears to be associated with making errors. If students respond too quickly, they may make mistakes. If, in Experiment 1, set completion was reinforced by escape from the laboratory session, then each frame is a conditioned negative reinforcer on a ratio schedule. Rapid responding (racing) produced reinforcement quickly, so racing became the dominant response pattern. A postfeedback delay would stop responses being immediately followed by a reinforcer (i.e., the next frame), so racing would not occur by adventitious reinforcement. Using delay in this manner is analogous to the technique of imposing delay in a multiple schedule. Experiment 2 assessed these two possibilities.
In addition, there seems to be two plausible means by which subjects working under Noncontingent Delay compensated for the time lost in imposed delays. First, because they made fewer errors during the first pass, they probably spent less time in review. Second, they may have increased their workrate at times when they were not locked out, in an attempt to minimise overall session length. Other studies also have reported that subjects work on the task during delays (e.g., Dannenbring, 1983), and increase their working pace during periods of task access (e.g., Stokes et al., 1988). Experiment 2 also assessed these possibilities.
Experiment 2

Experiment 1 found that Noncontingent Delay produced greater accuracy than No Delay. Two aspects of Noncontingent Delay that may have caused this were (a) an additional opportunity to inspect material, and (b) a delay that may have reduced racing. The present experiment was designed to assess both of these possibilities.

The present experiment provided a systematic replication of Experiment 1 with three main changes:

1. The following three conditions were employed in an alternating conditions design: (a) No Delay, (b) Noncontingent Delay, and (c) Noncontingent Delay with the screen blank during the delay period (Noncontingent Delay Blank Screen). If the racing notion is correct, and the presence of a delay per se was the important factor, then Noncontingent Delay and Noncontingent Delay Blank Screen would have similar effects and would both be superior to No Delay. Alternately, if additional study opportunity was the important factor, then Noncontingent Delay would lead to better accuracy than Noncontingent Delay Blank Screen and No Delay, which would have similar effects.

2. In the present experiment, four undergraduates (1 male, 3 female) who had no background in behaviour analysis were paid $10 for each session completed. Thus, the present experiment assessed the generality of the results obtained in Experiment 1 with subjects who differed in terms of background knowledge and motivation.

3. In Experiment 1, mean time per frame was longer under Noncontingent Delay than No Delay, but the extra time spent (about 4 s) was less than the 10-s imposed delay. This suggests that under Noncontingent Delay, (a) subjects spent less time reviewing incorrect responses (in reviews) because they made fewer mistakes, and/or (b) delays promote faster work during nondelay components of the frame. These propositions were difficult to assess because the efficiency measure employed in Experiment 1 (i.e., mean time per frame) did not distinguish between frames completed during the first pass through a set, and subsequent repetitions of incorrect items. Repeats may take less time because they are familiar, more time because they are difficult, or the same amount of time as first attempts. Although mean time per frame is an important measure of the overall efficiency of conditions, it was also important to gain a separate measure of the time required to complete the items during the first pass. The program used in the present experiment recorded time taken to answer all questions during the first pass, and total amount of time spent in enforced delays. The difference between these
represents actual time spent on task, or the rate at which subjects worked through materials.

**Method**

**Subjects**

One male and three female undergraduate students were paid volunteers in this experiment. H25 and H26 were female nursing students aged 18 and 19, respectively. H27 was a 19-year-old female psychology student, and H28 was an 18-year-old male enrolled in a Diploma of Tertiary Studies (see Appendix 2-2 for subject details). Each subject was paid $10 for each of the 15 sessions they completed (i.e., payment was not contingent on accuracy), plus an additional $25 bonus for not missing any scheduled sessions. None of the subjects had any experience with the subject matter used, and all were naive with respect to the aims and rationale of the present experiment.

**Apparatus**

**Design.** The present experiment was a partial replication of Experiment 1 in which the conditions No Delay and Noncontingent Delay did not change, but Contingent Delay was replaced by Noncontingent Delay Blank Screen (i.e., the question, response, and correct answer were removed from the screen during the delay period). Each subject received No Delay, Noncontingent Delay, and Noncontingent Delay Blank Screen in an alternating conditions design used in an identical fashion to Experiment 1. Hence, the sets used in baseline, No Delay, and Noncontingent Delay were the same as those used in Experiment 1, and the sets completed under Contingent Delay in Experiment 1 were completed under Noncontingent Delay Blank Screen in the present experiment (see Appendix 3-1).

**Procedure**

Before the first session subjects read instructions that were identical to those provided in Experiment 1, except that the three experimental conditions were described as (1) no delay between items, (2) a 10-second delay after each item, and (3) a 10-second delay plus a blank screen after each item. Subjects also used the same demonstration lesson used in Experiment 1, except that it was modified to present the new condition (Noncontingent Delay Blank Screen). Frame layout and screen sequence for No Delay and Noncontingent Delay were identical to Experiment 1. Events for Noncontingent Delay Blank Screen were identical to Noncontingent Delay except that the screen was blanked during the delay period. In Noncontingent Delay, after a subject had read an item and provided an answer, they corrected the response by pressing “C” or “I”; either of these responses initiated a 10-s delay in which the question, response, correct answer, Delay bar, and other
screen features were visible. In Noncontingent Delay Blank Screen, however, study materials (question, response, and correct answer) were removed from the screen during the delay. The other screen features (e.g., border, title, score box, Delay bar) remained visible. At the completion of the delay, the continuation message was presented in the normal fashion and, when a key was pressed, the next frame was initiated. After the study materials had been removed, they were not presented again in that frame.

Results

Accuracy (percentage of frames correct)

Figure 3-7 shows that, during baseline, accuracy was reasonably stable for two subjects (H27 and H28), more variable for one subject (H26), and ascending for the remaining subject (H25). For all subjects, scores became more variable during the alternating conditions phase. Visual inspection of the data suggests that Noncontingent Delay (filled circles) typically produced better accuracy than No Delay (unfilled circles; column 1) and Noncontingent Delay Blank Screen (filled triangles; column 2), and the latter two conditions did not differ systematically (column 3).

Table 3-4 shows that, across all subjects, Noncontingent Delay produced better accuracy than No Delay for 30 of the 38 sessions in which the conditions had different results, and this finding is significant ($p < .01$, sign test). This pattern also was evident in all four subjects. Noncontingent Delay also produced better accuracy than Noncontingent Delay Blank Screen for 27 of the 38 sessions, which also is significant, and each subject showed this pattern. In contrast, across all subjects, there was no systematic difference in accuracy between No Delay and Noncontingent Delay Blank Screen (19 of 38 sessions). Noncontingent Delay led to significantly greater accuracy than Noncontingent Delay Blank Screen and No Delay which did not differ. These results support the visual analysis and the extra-study proposition.

Figure 3-8 shows that, for each subject, Noncontingent Delay produced a greater percentage of correct frames than either of the other two conditions. Median results across subjects show that Noncontingent Delay produced 8% better accuracy than No Delay and 6% better accuracy than Noncontingent Delay Blank Screen, and that No Delay and Noncontingent Delay Blank Screen did not differ substantially. Hence, this form of analysis also supports the extra-study proposition.
Figure 3-7. For each subject, the percentage of frames correct per session (excluding reviewed items) for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Noncontingent Delay Blank Screen (NCB; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M”.
Table 3-4
Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Percentage of Frames Correct than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; NCB</th>
<th>ND &gt; NCB</th>
</tr>
</thead>
<tbody>
<tr>
<td>H25</td>
<td>6/9</td>
<td>8/9*</td>
<td>4/8</td>
</tr>
<tr>
<td>H26</td>
<td>9/10*</td>
<td>6/10</td>
<td>4/10</td>
</tr>
<tr>
<td>H27</td>
<td>6/10</td>
<td>6/10</td>
<td>6/10</td>
</tr>
<tr>
<td>H28</td>
<td>9/9**</td>
<td>7/9</td>
<td>5/10</td>
</tr>
<tr>
<td>Total</td>
<td>30/38**</td>
<td>27/38**</td>
<td>19/38</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and NCB = Noncontingent Delay Blank Screen.

*p < .05. **p < .01.

Figure 3-8. Percentage of first pass frames responded to correctly in baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Noncontingent Delay Blank Screen, NCB) for each subject. Median results across subjects are also shown.
Chapter 3

The finding of Experiment 1, that Noncontingent Delay leads to better accuracy than No Delay, was replicated in the present experiment. In addition, the magnitude of the effect was very similar in each case: 10% in Experiment 1, and 8% in the present experiment. Furthermore, because accuracy was not different under No Delay and Noncontingent Delay Blank Screen, the extra-study proposition is supported. That is, it is the presence of study materials during the delay, rather than the delay itself, which results in improved accuracy.

Reliability of self-scoring. For all subjects, self-scoring accuracy was greater than 95% (see Appendix 2-13), and there was no systematic bias towards any condition. Repeating the analyses presented in Table 3-4 after adjusting for errors in self-scoring did not alter the findings.

Review sets. As explained in Experiment 1, 3 of the 30 sets used in the experimental phase were reviews of previous sets. When these three sets were removed from the analysis of accuracy data, the pattern of significant results shown in Table 3-4 was unaltered.

Efficiency (mean time per frame)

In Experiment 1, mean time per frame (total time to complete set, including delays and repeats, divided by the number of frames in the first pass) did not differ consistently between Noncontingent Delay and No Delay, but, on average, the delay condition took approximately 4 s per frame longer. Similar results were expected in this experiment.

Figure 3-9 shows that mean time per frame was reasonably stable during baseline, and that, during the alternating conditions phase, it increased in variability for one subject (H26), and increased in variability and magnitude for another subject (H25). Visual inspection suggests that Noncontingent Delay Blank Screen took more time than Noncontingent Delay (column 2) and No Delay (column 3), which did not differ consistently (column 1).

Table 3-5 shows that, across all subjects, Noncontingent Delay had a greater mean time per frame than Noncontingent Delay Blank Screen in only 8 of the 39 sessions in which they differed. This is a significant result ($p < .01$, sign test). Similarly, mean time per frame was longer for No Delay than Noncontingent Delay Blank Screen in only 8 of the 39 sessions in which they differed. This also is significant ($p < .01$). There was no difference between Noncontingent Delay and No Delay (22 out of 39 sessions, $p > .05$). Furthermore, the pattern of results for individual subjects was consistent with the overall results.
Figure 3-9. For each subject, mean time per frame per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Noncontingent Delay Blank Screen (NCB; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M”.
Table 3-5

Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Mean Time per Frame than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; NCB</th>
<th>ND &gt; NCB</th>
</tr>
</thead>
<tbody>
<tr>
<td>H25</td>
<td>5/9</td>
<td>2/9</td>
<td>1/9*</td>
</tr>
<tr>
<td>H26</td>
<td>4/10</td>
<td>3/10</td>
<td>3/10</td>
</tr>
<tr>
<td>H27</td>
<td>7/10</td>
<td>2/10</td>
<td>2/10</td>
</tr>
<tr>
<td>H28</td>
<td>6/10</td>
<td>1/10*</td>
<td>2/10</td>
</tr>
<tr>
<td>Total</td>
<td>22/39</td>
<td>8/39**</td>
<td>8/39**</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and NCB = Noncontingent Delay Blank Screen.
*p < .05. **p < .01.

Figure 3-10. Mean time per frame during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Noncontingent Delay Blank Screen, NCB) for each subject. Median results across subjects are also shown. Weighted means have been used because varying set sizes and random assignment of conditions to sets resulted in slightly different numbers of frames in each experimental condition.
The data indicate that Noncontingent Delay Blank Screen consistently required more time per frame than the other two conditions which did not differ from each other. The lack of any consistent difference in overall efficiency between Noncontingent Delay and No Delay replicated the finding of Experiment 1.

Figure 3-10 shows mean time per frame across all sets. For each subject, Noncontingent Delay Blank Screen had a longer mean time per frame than the other conditions. Median results show that Noncontingent Delay Blank Screen took 12 s longer per frame than No Delay and 11 s longer per frame than Noncontingent Delay, and that there was little difference in the overall efficiency of No Delay and Noncontingent Delay. This is consistent with the results presented in Table 3-5. The similarity of the overall efficiency of Noncontingent Delay and No Delay is an important finding. In Experiment 1, subjects working in Noncontingent Delay compensated for about half of the time lost to delays but, in the present experiment, all time lost in imposed delays was recovered, and improved accuracy was achieved without any increase in the average time per frame.

First-pass Time

There is an important difference between the First-pass Time and mean time per frame measures. Mean time per frame provides an index of how long it takes to complete a frame when the inefficiencies of repeated items and imposed delays have been included. First-pass Time is an attempt to measure the work rate of subjects. It was calculated by dividing the overall time required to complete the first pass through a set by the number of frames in that set. Time spent in reviews and programmed delays are excluded from this measure. First-pass Time was employed to answer the question: During initial contact time with the materials, did subjects progress more quickly in Noncontingent Delay than in the other conditions?

Table 3-6 shows that, across all subjects, Noncontingent Delay had longer First-pass Times than No Delay for only 11 of the 37 sessions in which they differed and this is a significant result (p < .05, sign test). Similarly, Noncontingent Delay had longer First-pass Times than Noncontingent Delay Blank Screen for only 11 of the 37 sessions. This pattern, in which First-pass Times were consistently shorter under Noncontingent Delay than either of the other conditions, was evident for all four subjects. In contrast, across all subjects, there was no systematic difference in First-pass Time between No Delay and Noncontingent Delay Blank Screen (20 of 37 sessions, p > .05). In addition to having fewer frames to review, subjects work more quickly under Noncontingent Delay than No Delay. It is notable that this increase in work rate is counter to the racing proposition which proposed that imposed delays would increase accuracy by
slowing subjects: Instead, delays increase accuracy and increase workrate.

Table 3-6
Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater First-pass Time than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; NCB</th>
<th>ND &gt; NCB</th>
</tr>
</thead>
<tbody>
<tr>
<td>H25</td>
<td>3/9</td>
<td>2/9</td>
<td>4/9</td>
</tr>
<tr>
<td>H26</td>
<td>2/8</td>
<td>3/8</td>
<td>6/8</td>
</tr>
<tr>
<td>H27</td>
<td>3/10</td>
<td>4/10</td>
<td>5/10</td>
</tr>
<tr>
<td>H28</td>
<td>3/10</td>
<td>2/10</td>
<td>5/10</td>
</tr>
<tr>
<td>Total</td>
<td>11/37*</td>
<td>11/37*</td>
<td>20/37</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and NCB = Noncontingent Delay Blank Screen.
*p < .05.

Satisfaction

Figure 3-11 shows that, during baseline, ratings of satisfaction varied greatly; from high and stable (H25), to moderate and variable (H26), and also ascending (H27) and descending (H28), across sessions. During the alternating conditions phase there were no obvious changes in the level of satisfaction, but its variability increased. Visual inspection suggests that, for three subjects (H25, H26, and H28), the three conditions did not differ systematically in ratings of satisfaction. One subject (H27), however, consistently rated Noncontingent Delay Blank Screen lower in satisfaction than the other two conditions.

Table 3-7 shows that, across all subjects, there were no consistent differences in satisfaction between No Delay and either of the other conditions. The lack of difference between No Delay and Noncontingent Delay replicates the findings from Experiment 1. Across all subjects, Noncontingent Delay was rated as more satisfactory than Noncontingent Delay Blank Screen in 22 of the 28 sessions in which the ratings differed, which is a significant result (p < .01, sign test). Furthermore, this pattern occurred in each of the four subjects.
Figure 3-11. For each subject, satisfaction per session for the No Delay (ND; unfilled circles), Noncontingent Delay (NCD; filled circles), and Noncontingent Delay Blank Screen (NCB; filled triangles) conditions, for the last 5 sets in the baseline (Base) and the 10 alternating conditions sessions (AC). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked "M".
Table 3-7

Proportion of Sessions During the Alternating Conditions Phase in Which each Condition had a Greater Satisfaction than the Other Two Conditions

<table>
<thead>
<tr>
<th>Subject</th>
<th>NCD &gt; ND</th>
<th>NCD &gt; NCB</th>
<th>ND &gt; NCB</th>
</tr>
</thead>
<tbody>
<tr>
<td>H25</td>
<td>0/0</td>
<td>2/2</td>
<td>2/2</td>
</tr>
<tr>
<td>H26</td>
<td>4/6</td>
<td>4/6</td>
<td>4/7</td>
</tr>
<tr>
<td>H27</td>
<td>4/8</td>
<td>9/10*</td>
<td>9/10*</td>
</tr>
<tr>
<td>H28</td>
<td>7/9</td>
<td>7/10</td>
<td>3/9</td>
</tr>
<tr>
<td>Total</td>
<td>15/23</td>
<td>22/28**</td>
<td>18/28</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and NCB = Noncontingent Delay Blank Screen.

*p < .05. **p < .01.

Figure 3-12. Mean satisfaction per set during baseline (unfilled bars) and each of the three experimental conditions (No Delay, ND; Noncontingent Delay, NCD; and Noncontingent Delay Blank Screen, NCB) for each subject. Median results across subjects are also shown.
Satisfaction ratings show that, compared with No Delay, imposed delays (with or without study materials available) do not reduce consumer satisfaction. In addition, delays that included the study material were preferred over those that did not.

Figure 3-12 shows that satisfaction with the experimental conditions varied greatly across subjects. It is difficult to generalise from such variable data, but it is apparent that, for each subject, Noncontingent Delay had a moderate to high rating of satisfaction. Median satisfaction ratings across subjects show that Noncontingent Delay was rated 0.5 greater than No Delay and 1.0 greater than Noncontingent Delay Blank Screen. As was found in Experiment 1, there was no overall loss in satisfaction as a result of employing imposed delays.

Additional Findings

The percentage of errors made by subjects in the present experiment ranged from approximately 25% (H23) to 40% (H24), and resembled the findings of Experiment 1. Average time taken to complete an item also was similar to that in Experiment 1.

Discussion

Noncontingent Delay resulted in better accuracy than No Delay and Noncontingent Delay Blank Screen. This advantage had no cost in terms of mean time per frame or satisfaction, and First-pass Time showed that subjects worked faster under Noncontingent Delay. The advantage of a 10-s postfeedback delay is clear and replicable.

Noncontingent Delay Blank Screen was employed to test the racing and extra-study propositions. Compared with Noncontingent Delay, Noncontingent Delay Blank Screen resulted in fewer correct responses, more time per frame, and less satisfaction. Compared with No Delay, Noncontingent Delay Blank Screen did not differ in accuracy. Together, these findings support the proposition that postfeedback delays improve accuracy because they provide an opportunity for extra study. Accuracy is unaltered by postfeedback delays without study materials, so the racing proposition is not supported.

Negative Contingencies During Instruction

Imposing short-duration study periods improved accuracy with little or no overall time loss. Why is it then, that when subjects control their own pace (No Delay) they do not inspect materials for an additional period to gain the accompanying accuracy benefits? In No Delay, after a subject had scored an answer (i.e., typed "C" or "T"), the question, answer, and correct answer, were all displayed until a key was pressed, then the next question was displayed.
Consequently, after feedback was provided, subjects in any condition (except Noncontingent Delay Blank Screen) had access to all the relevant study material until they pressed a key. They could have paused at this point for 10 s, as they did in Noncontingent Delay, and obtained the same benefits. In other words, in No Delay, subjects were not obliged to proceed to the next frame immediately, and could have stopped and studied the material for as long as they wished. Previous researchers have predicted that subjects who have control over pacing are likely to lengthen the postfeedback interval as the task becomes more difficult (Bourne & Bunderson, 1963). Subjects in No Delay, however, did not spend more time studying materials.

Debriefing reports indicated that subjects' behaviour was controlled by negative rather than positive reinforcement. Although participation in the experiment was not intolerably aversive, and correct responding was enjoyable, activities outside the laboratory were more reinforcing; that is, subjects were motivated to escape the laboratory. Programmed instruction is assumed to be positively reinforcing because it is based on successive approximations whereby students complete many small, easy problems with a high degree of success. The broader contexts in which it is used (e.g., the classroom and laboratory), however, often operate on negative contingencies (e.g., complete the work before a deadline, or escape the session). Under such circumstances, it is reasonable that when subjects do not have to wait, they will not. In No Delay, subjects did not have to wait.

Subjects in the present experiments did not seem aware that by slowing their own progression they could improve their accuracy without an overall time loss. It is important to note, however, that the subjects' ability to monitor time, and estimate the proportion of correct and incorrect responses during a session, would have been very difficult due to the variation in set size, content, and difficulty. Nevertheless, the present results provide evidence that student self pacing, one of the basic tenets of programmed instruction and computer-assisted instruction may be suboptimal, and that brief, externally-imposed delays can increase accuracy.

Quicker Times due to Imposed Delays

In the time required to complete all items in a set correctly, subjects working under Noncontingent Delay can largely (Experiment 1), or completely (Experiment 2), regain the time lost due to programmed delays (mean time per frame). There are two main explanations for this. First, they do fewer frames: To finish a set, all questions must be completed correctly, and the most efficient way to do this is to respond correctly the first time and, thereby, minimise first-pass errors
and avoid repeating these items. Any manipulation which increases the number of 
questions answered correctly the first time will decrease session time, even if this 
manipulation is imposing a short delay between questions (Stokes et al., 1988, also 
found this with their keyboard-lock procedure). An important reason for this is that 
materials become more difficult as the set progresses, and later frames are based on 
earlier frames, so learning correct discriminations early is likely to increase the 
probability of getting later frames correct.

A second explanation for the small differences in efficiency (i.e., less than 
10 s per frame) across conditions is that subjects work more quickly under 
Noncontingent Delay (First-pass Time). Information obtained in debriefing 
suggests that delays increase escape motivation so that, under delay conditions, 
subjects worked more quickly to terminate the experimental session. This, 
however, is an inadequate account of the role of imposed delays upon work rate; 
subjects also experienced imposed delays under Noncontingent Delay Blank 
Screen, but their First-pass Times in that condition were not different from those in 
No Delay. Because increases in work rate occurred only in Noncontingent Delay, 
the availability of study materials appears to be necessary, and the role of escape 
motivation is unclear.

Perhaps the most complete explanation of the role of imposed delays 
combines both the above possibilities: Because subjects are motivated to finish the 
session, they work quickly when they have access to the task but, unlike 
Noncontingent Delay Blank Screen (which had less accurate and slower 
performance than Noncontingent Delay), only in Noncontingent Delay could 
subjects utilise the delays to study. Hence, in Noncontingent Delay subjects 
minimised errors, repeats, and time on task. Accuracy was good due to increased 
study, and this, in addition to escape motivation, led to a quicker work rate.

Remaining Issues

Experiments 1 and 2 raised some methodological and experimental issues that 
are the basis of the following experiments. These issues concern (a) whether the 
overall level of accuracy could be improved, and, if it could, whether the delay 
effect still occurred; (b) the precise means by which subjects reduced their First- 
pass Time under Noncontingent Delay; and (c) whether the gains observed during 
learning in Noncontingent Delay could be maintained on a subsequent, delayed test.

Accuracy. Although it has been argued that the error rates were not 
unacceptably high, further reduction of errors remains a laudable goal. One way in 
which this could be accomplished would be to pay subjects contingent on their 
accuracy (e.g., 15 cents per correct response, or $1 for each correct response on an
examination). Subjects in Experiment 1 were motivated to perform well because they were required to write research reports that were graded. In Experiment 2, however, there was no response-contingent payment: Subjects were paid $10 per session regardless of their accuracy and, during debriefing, some subjects commented that they were unconcerned if they made errors. Because there were similar error rates in Experiment 1 and Experiment 2, regardless of the subjects' vested interests, it was not clear what arrangement should be made in the next experiment. Subjects were not completing a course, so grades could not be made contingent on accuracy. Although financial incentives would probably work well, it was decided that a change in methodology may have had several advantages: Because some subjects had reported fatigue at the end of a session, shorter sessions may have removed the fatigue, increased attention, and, as a result, decreased errors. In addition, by not changing the payment contingencies, the next study remained comparable to Experiments 1 and 2. Hence, Experiment 3 employed shorter experimental sessions.

**First-pass Time.** Experiments 1 and 2 have both shown that subjects were able to compensate for the time lost due to enforced delays so that their overall time was not adversely affected. At present, the best explanation of how this was achieved involves a combination of fewer repeated items (accuracy) and a faster workrate (First-pass Time). It is still unclear, however, what behaviours contribute to this faster workrate: Subjects in Noncontingent Delay may have made up time in individual frames by responding to the question more quickly, or by correcting their response more quickly and continuing to the next frame. In Experiment 3 new timing measures were introduced to assist in resolving this issue.

**Retention.** Noncontingent postfeedback delays improve accuracy. Other important measures of academic behaviour are short and long-term retention, and to assess how delays influence retention Experiment 3 will introduce a pretest, posttest, and follow-up test.
CHAPTER 4: EFFECTS OF DELAYS ON WORKRATE AND RETENTION

Experiments 1 and 2 have shown that enforced delays increase accuracy with minimal, or no, loss in overall efficiency. This improved performance is not due to a delay per se, but occurs only when the question, response, and correct answer are present during the delay (Experiment 2). Several outstanding issues remain regarding the effects of noncontingent delays on subject behaviours, however, and these were addressed in the following studies. To accomplish this, several changes to both the outcome measures and the procedures were implemented. Because those changes are relevant to all the following experiments, they are explained below before the experiments are presented.

New Outcome Measures

Time Components

The measure of work rate obtained in the previous experiment (First-pass Time) did not provide sufficiently detailed information about time spent in each component of a frame to determine how time was saved under delays. To overcome this, and extend the analysis of how imposed delays influence subject behaviours, the controlling program was modified to record the following temporal components: (a) Question Response Time (QRT; time required to read the question and emit a response), and (b) Answer Question Time (AQT; time taken to score a response as correct or incorrect and move onto the next question). (The General Method details the sequence of on-screen events.) Although a measure of the interval between a response and answer (RAT; Response Answer Time) was possible, it is necessarily zero because answers were presented immediately following responses (i.e., feedback was instantaneous). Hence, this measure is not considered further. The First-pass Time is the sum of QRT and AQT, and each component was measured for each frame in the following experiments. At the end of each session the controlling program provided, on the summary output, the mean value for each variable for items in the first pass, review items, and all items (first pass plus review).

Retention

Experiments 1 and 2 established that Noncontingent Delay increases accuracy in learning sets, but they provided no information about retention. Retention studies are few in programmed instruction (Hartley, 1977), so it was important to measure the extent to which entering behaviour is increased by learning sets, how well improvements are maintained over time, and whether learning in the presence of
delays influences immediate and delayed retention. From Experiment 3 onwards, a pretest, an immediate posttest, and a one-month follow-up test were employed to address this. General procedures for test construction are described in the General Method, and study-specific details are provided with the relevant experiment.

Procedural Changes

Design

The alternating conditions design worked well in the first two experiments. Results were unambiguous within subjects, nonexperimental variables such as fatigue were controlled, and differences between conditions could be detected visually and statistically. The design could be improved, however. Presenting three sets per session resulted in long sessions, and subjects sometimes reported feeling fatigued. Fatigue may result in increased errors, particularly in the third set of a session, and although such a bias was counterbalanced across conditions, it is not ideal. It was decided to arrange two sessions per day (typically, one in the morning and one in the afternoon). In each session, subjects completed only one set under only one condition. In addition to decreasing fatigue problems, reducing the number of experimental conditions provided a focus on the main independent variable: In Experiments 3 and 4, No Delay and Noncontingent Delay were compared; Contingent Delay and Noncontingent Delay Blank Screen were not as effective in improving performance, and were not considered further.

This modified design enabled comparison of two conditions while controlling background variance. Because sessions were brief (one set only) and separated in time (by a minimum of 2 hrs), fatigue should have been reduced. In addition, because conditions were separated in time (one per session) and were clearly distinct (e.g., separated by a minimum of 2 hrs and correlated with discriminative stimuli), this should have reduced interactions between conditions.

Each of the following experiments employed a 10-session (5-day) baseline, a 20-session (10-day) alternating conditions phase, and a 10-session (5-day) reversal phase. Baseline and reversal phases consisted of only control conditions (No Delay), and the alternating conditions phase included the control condition and the experimental condition (Noncontingent Delay). This design maintained the number of experimental comparisons (i.e., 10) provided in Experiments 1 and 2, and also added a reversal phase. Reversal phases are employed to determine whether behaviour returns to baseline levels after experimental conditions have been removed, thereby demonstrating experimental control. By incorporating both an alternating conditions and a reversal phase this design provided assessments of
baseline behaviour before, during, and after the presence of the experimental condition.

**Review Sets**

In the first two experiments subjects completed each of the first 45 sets of the text (Holland & Skinner, 1961). Three of these sets (i.e., 17, 29, and 41) were reviews of material presented in the preceding section. Subjects completed these three sets in the normal fashion (under one of the experimental conditions), but because the items tested materials that had been studied in a previous set, under different conditions, performance in these sets may have been a function of practice, or initial exposure under another condition. In Experiments 1 and 2, analyses were conducted both with and without these sets, and their inclusion did not alter the findings. Nevertheless, because performance in review sets cannot be clearly identified with one experimental condition, they were omitted in future experiments. This is not optimal for learning purposes, but it was considered a necessary control procedure for the current experimental paradigm.

**Reviews**

Subjects repeated frames that they answered incorrectly. This review feature was employed because it is an important aspect of the ideal implementation of programmed instruction (Holland & Porter, 1961). Behaviour that occurs during a review is, however, qualitatively different from that which occurs in the first pass through materials because items have been seen previously, and they have been responded to in the context of the set (i.e., the context of sequentially-related items). Because of these differences, and because the majority of a subject's learning occurs during the first pass, data reported in future experiments will refer to the first pass unless otherwise specified. In addition, to increase comparability of no-delay and noncontingent-delay conditions, delays were not employed in the review of Noncontingent Delay and, in this way, the reviews of each condition were standardised. This was not the case in Experiments 1 and 2 where the experimental contingencies were employed in both the first pass and the review.

**Experiment 3**

The general aim of the present experiment was to refine the experimental procedure so that the maximum potential of imposed delays could be determined. More specific aims were: (a) to introduce methodological changes designed to reduce subject fatigue, reduce overall error rates, and facilitate comparison of experimental conditions; (b) to replicate the major findings of Experiments 1 and 2 with the new procedures; (c) to measure subject response latencies at different
positions within a frame to determine the precise means by which subjects reduce First-passage Time under Noncontingent Delay; and (d) to consider the longer-term effects of instruction using retention tests.

It was expected that, overall, subjects would be more accurate than in previous experiments; that imposed delays would again improve accuracy, and that this would be reflected in better retention of materials learned under delay conditions. Of particular interest was whether any advantage of delays on the posttest would be increased, decreased, or remain the same on the follow-up test. It was expected that imposed delays would again decrease First-passage Time, and the component measures would help identify the behaviours leading to this effect.

Method

Subjects

Three student volunteers were paid for their participation in this experiment: H35 was a 22-year-old, female, masters-level chemistry student; H36 was a 21-year-old, female, undergraduate biology student; and H37 was a 19-year-old, male, undergraduate mathematics student (see Appendix 2-2 for subject details). Subjects received $5 for each session completed, plus a bonus payment of $50 if they did not miss any experimental sessions. If subjects missed one or more sessions, they did not receive the bonus and, additionally, lost $15 for each session missed. Subjects were recruited in the same way as described previously. All subjects were naive with respect to the aims and rationale of the present experiment, and had no experience with the subject matter used.

Apparatus

Design. The design of this experiment was changed from that used in the first two experiments in two ways: (a) in addition to a baseline and alternating conditions phase, a reversal phase was included; and (b) two conditions were used instead of three. These conditions were (a) No delay between frames (No Delay), and (b) A 10-s delay after each frame (Noncontingent Delay). Two sessions were completed each day, and only one set was completed in each session; each set (or session) took approximately 20 min to complete and sessions were separated by a period of at least 2 hrs. In Sessions 1 to 10 (baseline), all sets were presented with No Delay. In Sessions 11 to 30 (alternating conditions phase), sets were presented with either No Delay or Noncontingent Delay, with the order being determined by an unconstrained random sequence (see Appendix 4-1). Sessions 31 to 40 provided return to baseline conditions, and all sets were presented with No Delay. The number of frames completed in each condition were as follows: 352, baseline; 420 No Delay; 476, Noncontingent Delay; and 311 reversal.
Retention. A single test was constructed for use as a pretest, posttest, and follow-up test, and permitted the direct comparison of results from each testing session. The test needed to be sufficiently large to be comprehensive and discourage rote memorisation of items (particularly between the posttest and follow-up test), but not so large as to exhaust subjects or make them too anxious in the pretest (i.e., by performing poorly on many items). Fifty items was considered a reasonable test size. To select the items for the test, a pool of items was constructed from the review sets included in the text (Holland & Skinner, 1961). This pool included only those items that were relevant to the portion of the text that was presented to subjects (i.e., the first 40 nonreview sets; see General Method for more detail). Fifty items were randomly selected without replacement from a pool of 96 eligible items, so that baseline, No Delay, Noncontingent Delay, and reversal were represented by 14, 15, 12, and 9 items, respectively.

Procedure. On the first day of the experiment, subjects read the experimental instructions (which were slightly modified to reflect the reduced number of conditions and the addition of tests; see Appendix 2-4), interacted with the demonstration program (modified for the 2-condition design), signed a consent form (see Appendix 2-7), and completed the pretest. On the second day they began the series of learning sets (which took 20 days to complete) and, on the day following this period, they completed the posttest. After one month, subjects completed the follow-up test and were debriefed.

Results

Accuracy (percentage of frames correct)

Figure 4-1 shows that, during baseline, accuracy was high (typically over 80% of items in each set answered correctly) and reasonably stable for each subject. During the alternating conditions phase, accuracy under Noncontingent Delay (filled circles) was generally high and stable, whereas accuracy under No Delay (unfilled circles) was more variable. In No Delay, accuracy appeared to decline over sets for each subject (with the exception of the third last set of H35 and H37), although for H36 the decline is only apparent in the final sessions of this phase. During the reversal phase, the level and variability of accuracy was similar to baseline for two subjects (H35 and H37), and the accuracy of the remaining subject (H36) showed a recovery to baseline levels.
Figure 4-1. For each subject, percentage of frames correct per session (excluding reviewed items), for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison.
If the alternating conditions phase is considered alone, it appears that accuracy under Noncontingent Delay was typically higher and less variable than accuracy under No Delay. Once again, imposed delays appeared to improve accuracy.

Consistency of Noncontingent Delay effects is shown by the proportion of sessions in which it resulted in a greater percentage of frames correct than No Delay. For all subjects, the percentage of frames correct was higher for Noncontingent Delay than No Delay for 22 of the 29 comparisons (for those sets for which scores differed), which is significant by a sign test ($p < .01$). This pattern of better accuracy under Noncontingent Delay was evident for all three subjects (8/10 for H35; 7/10 for H36; and 7/9 for H37). In addition, it replicates the consistent advantage of Noncontingent Delay found in Experiments 1 and 2.

Figure 4-2 shows, for each subject and condition, the percentage of frames answered correctly during the alternating conditions phase. It is noteworthy that it does not present median results; none of the bargraphs in the present, or subsequent, experiments do. Although similar bargraphs in Experiments 1 and 2 included medians, subsequent experiments each employed only three subjects, and median results are relatively easy to detect from these. Removing medians avoids replicating data and detracting from the individual's results.

Figure 4-2 shows that, for each subject, Noncontingent Delay produced more correct responses than No Delay. The magnitude of this advantage is 10%, 5%, and 8% for H35, H36, and H37, respectively, and is very similar to the advantage of Noncontingent Delay in Experiments 1 and 2 (a 9.7% and 8.4% median increase in accuracy, respectively). The magnitude of the delay effect was maintained in the present experiment even though the experimental design changed considerably from that used in previous experiments. Moreover, in the present experiment, this improvement due to delays occurred when accuracy levels were quite high (approximately 80% in the alternating conditions phase). Hence, high levels of accuracy do not appear to place a ceiling constraint on the experimental effect.

Reliability of self-scoring. For all subjects, self-scoring accuracy was greater than 98% (see Appendix 2-13), and there was no systematic bias towards either condition. Repeating the analysis after adjusting for self-scoring errors did not change the pattern of results or the outcome.
Figure 4-2. Percentage of frames correct in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject.

Efficiency (mean time per frame)

Figure 4-3 shows the mean time required to complete a frame in each session (i.e., time spent in first pass and review plus any imposed delays, divided by the number of frames in the first pass). For each subject, the mean time per frame under No Delay maintained a consistent level (i.e., it did not increase or decrease) across sessions, although there was a moderate amount of variability. Visual inspection indicates that mean time per frame was less variable under Noncontingent Delay, and also was consistently longer under this condition. Binomial comparisons support the notion that mean time per frame is typically longer under imposed delays: Across all subjects, the proportion of sessions in which mean time per frame was longer under Noncontingent Delay than No Delay was 24/30 which is significant ($p < .01$, sign test), and this pattern was consistent across all subjects (7/10 for H35; 7/10 for H36; and 10/10 for H37). This finding of consistently lower efficiency under Noncontingent Delay is contrary to the findings of the previous two studies in which there has been no consistent loss in efficiency (binomial comparisons) due to delays.
Figure 4-3. For each subject, the mean time per frame in each session for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles).
Figure 4-4 shows that for each subject, and across all frames in the alternating conditions phase, the mean time per frame was greater under Noncontingent Delay than No Delay. For two subjects (H35 and H36) the magnitude of the difference across conditions was approximately 5 s which indicates that these subjects recouped some of the time lost during the enforced delay. For the remaining subject (H37), however, the difference in mean time per frame of the no-delay and noncontingent-delay conditions was approximately 9 s (i.e., approximately the duration of the imposed delay), so he saved only a small amount of time.

![Graph showing mean time per frame for different conditions](image)

**Figure 4-4.** Mean time per frame in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject.

**First-pass Time and Time Components**

There are three considerations when interpreting the temporal measures: (a) First-pass Time is the numerical sum of the other time components (QRT, Question Response Time; AQT, Answer Question Time); (b) QRT represents time required to read the problem and emit a response, and AQT represents time taken to score a response as correct or incorrect and move onto the next item; and (c) When the computer recorded the AQT it included the 10-s imposed delay for Noncontingent Delay, so this amount was removed from the data before any subsequent analyses were conducted.

Table 4-1 shows, for each subject, and across all subjects, the consistency of the effect of Noncontingent Delay on First-pass Time and each time component of a
frame. Overall, there was no consistent reduction in First-pass Time in Noncontingent Delay. This finding is counter to that of Experiment 2 in which Noncontingent Delay did reduce First-pass Time overall and for each subject. The inconsistent outcomes across experiments cast doubt on the notion that subjects work faster under imposed delays due to increased escape motivation. The overall results for the time components showed that Noncontingent Delay resulted in a statistically significant reduction in QRT (sign test; \( p < .05 \)), but AQT was not different across conditions.

Table 4-1
Proportion of Sessions in Which Noncontingent Delay had a Greater \(^a\) First-pass Time, Question Response Time, and Answer Question Time than No Delay

<table>
<thead>
<tr>
<th>Subject</th>
<th>First-pass Time</th>
<th>QRT</th>
<th>AQT(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H35</td>
<td>2/10</td>
<td>1/10*</td>
<td>3/9</td>
</tr>
<tr>
<td>H36</td>
<td>4/10</td>
<td>4/10</td>
<td>4/10</td>
</tr>
<tr>
<td>H37</td>
<td>6/10</td>
<td>5/10</td>
<td>9/10*</td>
</tr>
<tr>
<td>Total</td>
<td>12/30</td>
<td>10/30*</td>
<td>16/29</td>
</tr>
</tbody>
</table>

\(^a\)The mean value from each set was used for comparison. \(^b\) AQT values for the Noncontingent Delay had 10 s subtracted before comparison to adjust for imposed delays.

\(* p < .05.\)

Since the overall findings do not necessarily represent the pattern of responding of individual subjects, their idiosyncratic response patterns are discussed separately. For H35, First-pass Time was greater under Noncontingent Delay than No Delay in only 2 of the 10 comparisons, indicating that this subject worked more quickly under Noncontingent Delay. Consistent with this finding, each of the component time measures (QRT and AQT) also was typically smaller under Noncontingent Delay (1/10 and 3/9, respectively). These results indicate that, under conditions of imposed delays, the work rate is increased at each stage of the frame, and may occur more during QRT; perhaps because it is a longer interval and more likely to enable substantial time savings. H36 showed a trend towards quicker performance under Noncontingent Delay for First-pass Time and for each of the
time components, but these results were not different from chance. Subject H37 showed no systematic differences, across conditions, in First-pass Time or QRT, and this is in accordance with the earlier finding that he made little or no compensation for the time lost due to imposed delays (efficiency data). This subject did, however, have a consistently longer AQT under Noncontingent Delay than No Delay (9/10 comparisons), and there are two possible interpretations of this finding: Under enforced delays the subject spends more time either (a) scoring his responses, and/or (b) responding to the prompt to initiate the next frame. These alternatives have quite different implications, the former suggesting careful attention to the task (attending the stimulus, response, and feedback) whereas the latter implies distraction from the task (assuming subjects initiate the next frame as soon as possible). This difficulty in interpreting the finding is overcome in subsequent experiments by recording separately the time to correct responses and the time to continue to the next frame.

There were no reliable differences in First-pass Time (unlike Experiment 2) or its temporal components, so these measures have not been helpful in determining which behaviours, if any, have their speed influenced by imposed delays. Possible reasons for the inconsistent outcomes will be discussed shortly, along with a solution that provides the basis for Experiment 4.

Satisfaction

Figure 4-5 shows the satisfaction ratings of each subject in each session. Overall response patterns are difficult to describe because each subject's results differ considerably from the others. For one subject (H35), satisfaction with No Delay was high and stable during baseline, high and somewhat variable in the alternating conditions phase, and high and relatively stable in the reversal phase. Ratings of Noncontingent Delay were generally high with moderate variability. Visual inference suggests no clear preference for either condition. For the second subject (H36), ratings of satisfaction with No Delay were highly variable and showed a general decline across the duration of the experiment. Ratings of Noncontingent Delay also were variable and do not appear to be consistently different, in level, from No Delay. The remaining subject (H37) provided high and stable ratings of satisfaction under No Delay throughout the experiment (with two exceptions during baseline). Noncontingent Delay was initially rated low, but satisfaction with this condition increased throughout the alternating conditions phase and, by the final session in this phase, was rated equal to No Delay.
Figure 4-5. For each subject, satisfaction per session for No Dclay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles).
This subject had a preference for No Delay which is visually apparent, but the magnitude of this preference diminished during the experimental phase.

Visual inference is supported by binomial comparisons. The number of sessions in which Noncontingent Delay had a higher satisfaction rating than No Delay was 3/8 for H35 and 5/9 for H36, indicating no consistent preference for either condition. For H37, in none of the sessions in which ratings differed was Noncontingent Delay rated more highly than No Delay (0/9), showing a clear preference for No Delay and also accounting for the significant overall preference for No Delay (8/26 comparisons; p < .05, sign test).

Figure 4-6 shows that, for each subject, mean satisfaction per session is higher for No Delay than Noncontingent Delay. For H35, both experimental conditions were rated high and the difference between their ratings was 0.2; for H36, both conditions were rated low and the difference between them was 0.3; and for H37, No Delay was rated high and 4.3 greater than Noncontingent Delay. Like the consistency data, mean satisfaction ratings show that there was little difference in the preferences of two subjects, whereas one subject clearly preferred No Delay.

![Bar Chart](image)

**Figure 4-6.** Mean satisfaction ratings in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject.
It is unclear whether the session-by-session ratings of satisfaction are helpful or have construct validity. Figures 4-5 and 4-6 show that satisfaction ratings fluctuate greatly both within and between subjects and, in the next section, it will be shown that this variability is in contrast to the distinct preferences reported by subjects during debriefing.

**Subject Preferences**

Debriefing reports showed that, although subjects did not dislike either condition, they all expressed a preference for No Delay, and the reason for this was straightforward: Delays prolong the experimental session, and they do so regardless of accuracy (i.e., they were not contingent on response correctness). When a response was correct, subjects were kept waiting for no apparent reason; when a response was incorrect, subjects only required the study opportunity for a few seconds, and the remaining period of delay was, again, an unwanted (and seemingly unnecessary) wait. This is further anecdotal evidence that subjects were motivated to escape the experimental session, and this proposition was investigated further in Experiment 4. Subjects suggested that 5 s would be a more appropriate delay length because that was the amount of time required to resolve errors or clarify problems. Indeed, subjects acknowledged that, for those reasons, Noncontingent Delay was sometimes beneficial to learning.

One way to increase consumer satisfaction with Noncontingent Delay (if this is necessary) may be to make consumers aware that they perform better in this condition. This awareness does not appear to occur incidentally: When explaining her preference for No Delay, H35 suggested that it is because she performed better in this condition - this is definitely not the case (see Figure 4-2). To increase awareness of relative performance, feedback such as a running mean percentage correct for each condition could be maintained on the screen. There was an on-screen scorer at all times, but this only indicated performance in the current session, and did not compare conditions. Of course, if comparative performance feedback was built in to the present experiments, it would confound the experimental variable.

**General opinions vs. session ratings.** The session-by-session satisfaction ratings do not provide a straightforward measure of satisfaction with the conditions. Ratings varied greatly both within and across subjects (Figure 4-5), and relative ratings of Noncontingent Delay and No Delay do not correspond well with subjects' unambiguous general opinions of those conditions. Two subjects (H35, H36) expressed a preference for No Delay that was in contrast to the overlapping session-by-session ratings (see Figure 4-5) and the similarity of mean satisfaction
ratings (see Figure 4-6). Only the remaining subject (H37) showed a correspondence between his general opinion (preferred No Delay) and his satisfaction ratings.

Ratings of satisfaction may fluctuate widely and lack correspondence to general opinions because they are influenced not by the experimental condition alone, but by performance, fatigue, escape motivation, hunger, and other uncontrolled factors that operate within each session. Because it is unclear what determines satisfaction ratings, and, therefore, what they represent precisely, opinions expressed during debriefing may be a better index of consumer satisfaction with experimental conditions. If there is further discordance in these measures in subsequent experiments, only subjects' general opinions will be presented.

Additional Findings

Figure 4-2 shows that subjects responded incorrectly to approximately 20% and 15% of no-delay items and noncontingent-delay items, respectively. Error rates in each of these conditions are considerably lower than the respective values in Experiments 1 and 2 (see Figures 3-2 and 3-8), and, particularly those of Noncontingent Delay, more closely resemble the error rates typically expected from well-programmed materials. This overall increase in accuracy may be due to the methodological changes designed to reduce fatigue, or due to subject selection. More importantly, it indicates that the procedural changes employed in the present experiment have not reduced accuracy levels, or removed the benefit of delays.

In addition to making few errors overall, subjects in the present experiment also worked quickly. Time required by subjects to complete sets was compared to the estimated times provided in the text (Holland & Skinner, 1961). One subject (H35) typically completed materials at a similar or faster rate than those published, another (H37) typically worked at the same pace, or slightly slower, than the published times, and the remaining subject (H36) responded more slowly. It is worth reiterating that the published times are regarded by the authors as guides only, and are shorter than most students can accomplish. The present subjects performed well, and although there are various differences between the present implementation of the program and that of Holland & Skinner (such as performance contingencies, time pressures and deadlines, and student ability), it is important that the present subjects' high level of performance cannot be attributed to taking an extraordinarily long time with the materials.
Retention

Figure 4-7 shows the percentage of correct responses provided by each subject on each test. This figure does not distinguish between experimental conditions, but instead enables a comparison of overall pretest, posttest, and follow-up test results. There is a clear trend in the data: all subjects performed substantially better in the posttest than the pretest, indicating that entering behaviour was substantially improved by the learning material. Although it is possible that events such as practice or passage of time may account for some improvement, it is reasonable to assume that gains of this magnitude are primarily due to exposure to learning materials. Furthermore, the improvement was largely maintained in the one-month follow-up test: H35 showed a slightly improved recall (4%), and H36 and H37 showed decrements of approximately 10%.

![Bar chart showing percentage correct for H35, H36, and H37 across pretest, posttest, and follow-up tests.]

**Figure 4-7.** For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The same 50-item test was used on each testing occasion.)

Figure 4-8 shows, for each subject, the percentage of correct responses in each test (pretest, posttest, and follow-up test) for each of the four conditions of the learning phase (Base, ND, NCD, and Base2). There is a clear general trend: in both the posttest and follow-up test there is a general pattern of decline in the proportion of correct items across phases: retention was best for materials studied during baseline, worst for materials studied in reversal, and intermediate for material studied in the alternating conditions phase. This is reasonable because materials become more complex as the program progresses.
Figure 4-8. For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). (Baseline, No Delay, Noncontingent Delay, and reversal were represented by 14, 15, 12, and 9 items, respectively.)
In addition to general trends, there were clear differences between experimental conditions. In the posttest, all subjects showed better retention of materials studied under Noncontingent Delay than under No Delay. The advantage due to delay was large for 2 subjects (22% for H35 and 23% for H37) and small (5%) for the remaining subject (H36).

In the follow-up test, the advantage for Noncontingent Delay increased substantially for 2 subjects (by 20% for H36 and 13% for H37) because, relative to the posttest, retention of materials studied under No Delay decreased but materials studied under Noncontingent Delay were well maintained. For the remaining subject (H35), however, the recall advantage was reversed and no-delay materials were recalled better at follow-up. Like the other subjects, H35 showed good retention of Noncontingent Delay material at follow-up but also showed an unexpected and large increase in recall of no-delay materials. It is unclear why this subject improved recall of the no-delay material from the posttest to the follow-up test, but some possibilities are considered in the Discussion. Overall, material studied under Noncontingent Delay was recalled better and maintained well.

Summary

During learning sets, all subjects made more correct responses under Noncontingent Delay than No Delay, and the magnitude of the difference was comparable to that found in the previous two experiments. In the present experiment, however, that difference occurred despite an overall increase in the number of correct responses being made. In accordance with performance in the learning sets, immediate and delayed retention was typically greater for materials studied under Noncontingent Delay. Sets in which a delay was imposed after each frame took longer to complete, and although some time savings were apparent under this condition, faster workrates did not occur reliably. It became apparent that satisfaction ratings do not necessarily correspond with general impressions of the conditions, and the latter may be better indicators of subject preferences.

Discussion

Accuracy

The present experiment replicated the finding that subjects make fewer errors under Noncontingent Delay than No Delay. The magnitude of the experimental effect was very similar to that of Experiments 1 and 2: In those studies between 3.6% and 13.7% improvement, across subjects, resulted from the use of Noncontingent Delay; in the present experiment, the range of improvement was from 4.5% to 9.5%. Furthermore, the mean error rate of the present experiment
(7.3%) is very similar to the previous medians (9.7% and 8.4% for Experiments 1 and 2, respectively).

Two important implications arise from the replication of the delay effect in the present experiment:

1. **Based on the findings of Experiment 1, it was suggested that when accuracy in No Delay is high, Noncontingent Delay only shows small improvements (see H22). This is not supported by the present results because subjects who had high accuracy in No Delay also showed substantial differences between conditions. Hence, this is strong evidence of the control exerted by the delay variable.**

2. **The present experiment replicated the delay effect using a different design from previous experiments, and this enables condition-to-set assignment to be ruled out as a potential confound. In Experiments 1 and 2, assignment of conditions (No Delay and Noncontingent Delay) to sets was identical (Appendix 3-1), and it was possible that improved accuracy under Noncontingent Delay was due to fortuitous assignment of this condition to easier sets. The present experiment rules out this possibility because condition-to-set assignment (Appendix 4-1) was different from that used previously: Only 4 of the 10 sets that used No Delay in the experimental phase of Experiments 1 and 2, also used this condition in Experiment 3. Similarly, only 4 of the 10 sets that used Noncontingent Delay in Experiments 1 and 2 also used this condition in Experiment 3. Remaining sets either were assigned to another condition, or used in a different phase and were not used in the analyses. In general, delays improve accuracy regardless of the sets to which they are assigned.**

Why were subjects in the present experiment substantially more accurate than previous subjects? Three factors may account for this:

1. **The present experiment used 1-set sessions rather than 3-set sessions, so these subjects may have been less fatigued and more attentive to stimulus materials.**

2. **Subjects may have been more clever. This notion is difficult to support, however, because subjects’ (from all three experiments) academic records were checked, and found to be poorly related to their performance in learning sets.**

3. **During debriefing, each subject expressed great concern about making large numbers of errors and thereby appearing unintelligent. This social pressure was not as evident in Experiments 1 and 2 where some subjects reported that they were unconcerned about making errors. It is not clear whether some aspect of the present design or procedure contributed to this pressure, or whether it was a sample-specific attribute. This concern for their own performance resembles**
realistic (academic) settings and, if this is the case, it increases the external validity of the present experimental paradigm.

Finally, it is noteworthy that the overall increase in accuracy in the present experiment, relative to the previous experiments, occurred in spite of two procedural changes which would be expected to promote errors. First, in Experiments 1 and 2, items which were answered incorrectly in the first pass were repeated in a review, and these repeated items also were accompanied by the experimental condition (which was often a delay). In the present experiment, however, delays were not used in the review of Noncontingent Delay and this would presumably have made first-pass errors less aversive and, hence, more likely to occur. Second, subjects in the present experiment did not have the advantage of intermittent review sets (i.e., Sets 17, 29, and 41 of Holland & Skinner's text) which were omitted for purposes of experimental control.

Activity During Delays

Experiment 2 demonstrated that noncontingent delays are only advantageous when accompanied by study materials. The extra-study explanation was spontaneously expressed by subjects in the present experiment. During debriefing, two subjects (H35 and H37) reported that they used the 10-s delay period to study stimulus materials, whereas the remaining subject (H36) claimed that she purposefully avoided studying the materials during this interval. The two subjects who claimed that they used the extra-study opportunity benefited more from the experimental condition (see Figure 4-2) than the other subject. Notwithstanding this, H36 also showed better accuracy (consistency across sessions and magnitude of effect) under Noncontingent Delay without using an explicit study strategy, which is evidence that there is also an incidental advantage to working under this condition. Importantly, subjects could describe two different uses of the delay period: Sometimes they studied the question, their response, and the correct-answer feedback in order to modify an incorrect response or consolidate a correct response (the latter practice may explain why noncontingent delays increased accuracy beyond response-contingent delays in Experiment 1); at other times they studied the correct-answer feedback alone in order to reproduce it later in the review. Both strategies were aimed towards reducing time spent in review, and this is further evidence that subjects are motivated by escape from the experimental session. The escape motivation hypothesis will be discussed in more detail below.

Retention

The present experiment demonstrated that noncontingent delays increase retention. There was a clear posttest advantage for material learned under
Noncontingent Delay. Furthermore, this advantage increased over time (to the one-month follow-up) for two reasons: (a) noncontingent-delay materials were well maintained, and (b) recall of no-delay materials decreased. One exception to this general pattern was the substantial (20%) increase, for H35, in recall of no-delay material from the posttest to the follow-up test. This was an unexpected outcome which may indicate problems with the assessment tool. If this subject was merely inattentive on some no-delay items in the posttest, and corrected this in the follow-up test, then the results are consistent: Materials studied under Noncontingent Delay are well recalled and well maintained, whereas materials studied under No Delay are recalled less well and are not well maintained. An alternative explanation is based on variability in subjects’ responses and the criteria for scoring responses as correct. A subject’s constructed response to any one test item may differ across testing occasions, so that a response given on one occasion (posttest) may be determined by the scorer not to be among a range of correct alternatives; on another occasion (follow-up test), however, the subject may provide a slightly different response that is considered correct by the marker. If this occurred, test scores could increase over time. If this is the case, it is unclear how large a retention advantage for Noncontingent Delay must be to be regarded substantial and reliable.

Notwithstanding the follow-up test results of H35, there is a consistent relation between subjects’ reported behaviour during learning sets and differences between the conditions in both learning sets and tests. When delays were present during learning sets, those subjects (H35 and H37) who claimed that they used this opportunity to study showed larger accuracy differences across conditions than the subject (H36) who claimed not to study. Presumably because they understood the material better, H35 and H37 also showed greater posttest differences between conditions than did H36. This makes sense. If delays provide the opportunity to study, then more study should lead to greater accuracy, and this should result in better retention. (Except for the unusual results of H35, the same pattern occurred at follow-up.)

There is an experimental confound that is an obstacle to relating test results to the experimental condition: Noncontingent-delay materials may have been recalled better because they were studied under a “novel” condition. That is, subjects may have paid more attention to noncontingent-delay materials because they were a change from the no-delay conditions of baseline. To properly determine whether the delay or novelty component of Noncontingent Delay accounts for the test results requires direct experimental investigation, but indirect evidence can be drawn from Experiments 1 and 2. In each of those experiments, there were two novel
conditions introduced during the alternating conditions phase, but only Noncontingent Delay improved accuracy. This indicates that this condition has an influence on learning other than novelty, and it is not unreasonable to expect this influence to extend to recall.

**Motivation**

Although the accuracy and retention data are clear and reasonably consistent, the timing measures have been less consistent. In Experiments 1 and 2, efficiency (mean time per frame) was similar regardless of the presence or absence of delays. It was logical that, although external pacing with delays (Noncontingent Delay) decreased efficiency, self pacing (No Delay) was also inefficient because it typically led to more errors and, hence, more time lost repeating items. Perhaps more interesting was the possibility that subjects work faster during Noncontingent Delay as a means of compensating for time lost in imposed delays.

Unfortunately, the data from Experiment 3 do not increase our understanding of the influence of imposing delays upon workrate (First-pass Time) or overall efficiency, because the two main timing measures have provided results that are inconsistent with previous experiments. Although Experiments 1 and 2 showed that there was no consistent difference between No Delay and Noncontingent Delay in mean time per frame (which was evidence that external pacing did not reduce overall efficiency), the present experiment found that Noncontingent Delay did consistently increase mean time per frame over No Delay. Another unexpected finding was that, unlike Experiment 2, the present experiment found that First-pass Time was similar in No Delay and Noncontingent Delay. This indicates that imposed delays do not influence subjects’ workrate, and places in doubt the notion that imposed delays increase escape motivation. The motivation issue is worth pursuing because most subjects have reported that it was important to complete the session as quickly as possible, and it is reasonable to assume that delays increase motivation to escape.

The inconsistencies in timing data may be due to the uncontrolled influence of review size. In Experiments 1, 2, and 3, review size was not controlled because it was assumed that repetition of incorrect items had solely an educative role which had an equal influence on performance in all conditions, and which would not interfere with experimental manipulations or confound conclusions. However, the review may have other functions: The procedure of repeating incorrect items may be conceptualised as negative reinforcement of correct responses (or punishment of incorrect responses) because the size of a review is incremented by one item contingent on each incorrect response. Some subjects described reviews as an
index of cleverness: Although a few mistakes were considered acceptable, large reviews meant they were unintelligent. If subjects were motivated to minimise errors, then reviews may increase accuracy; whether reviews would have a different effect across pacing conditions, however, is unclear. Reviews may also influence workrate. For example, if subjects were aware that they made more errors under No Delay, and would have to complete many time-consuming repeats, it is possible that they may have increased their workrate under this condition. Whether reviews have any such effects needs to be assessed empirically. Hence, the following experiments were designed to address two main issues: (a) What are the effects of externally-imposed delays on performance when review size is controlled? and (b) What are the effects of review size on performance (in particular, the cost in efficiency of repeating items)?
Experiment 4

Experiment 4 investigated the influence of externally-imposed delays on performance with review size controlled across conditions. It was important to replicate accuracy and retention advantages associated with Noncontingent Delay found in Experiment 3, and also to assess the influence delays have on workrate and overall efficiency when other time influences are removed.

Escape Motivation in Previous Studies

There is good evidence that subjects attempt to minimise time on task in programmed instruction. In an investigation of subject eye movements during completion of programmed materials (Doran & Holland, 1971), it was found that students first fixate near the response blank and then appear to search the item for key words. (Subjects in the present experiments also claim to having used this strategy for efficiency.) This indicates that students will readily skip information that is not essential for obtaining the correct answer, probably because working out the correct answer quickly is more important than reading all of the material. Subjects in that experiment undoubtedly had high escape motivation; to enable precise calibration of the eye movement equipment they were constrained in a dentist chair with a bite bar.

There also is ample evidence that delays influence efficiency in computer-based tasks. When computing tasks were commonly conducted on time-share systems, users often experienced delays of unpredictable onset and duration. It was forecast that such delays would impede progress and be detrimental to accuracy. When researchers systematically investigated user behaviour under different conditions of system-response delays, however, they found that subjects who experienced delays became more efficient in numerous elements of the task repertoire so that they fully utilised the contact time that was available (e.g., Dannenbring, 1983; Grossberg et al., 1976). Similar increases in efficiency also have been demonstrated in more recent investigations of delays in CBI (e.g., Stokes et al., 1988).

Escape Motivation in the Present Studies

The present studies have used First-pass Time as a measure of subjects' motivation to complete the experimental session as quickly as possible. Although there has been ample anecdotal evidence that delays increase escape motivation (and, hence, reduce the inefficiencies of imposing delays), the experimental findings have not been entirely consistent. In Experiment 2, First-pass Time indicated that all subjects consistently worked more quickly under delay conditions than no-delay conditions, but in Experiment 3 this trend only occurred
for one subject and there was no significant overall effect. It has been argued that this inconsistency may have arisen due to the uncontrolled influence of review size upon workrate.

In Experiments 1, 2, and 3, any items answered incorrectly during the first pass were repeated in a review. Hence, reviews increased the number of items per session and the total session length. But the influence of reviews may not only be increased time at the end of a set; they may influence workrate throughout a set. Just as imposed delays may cause compensatory increases in workrate when subjects are not locked out, so might reviews. Because both delays and reviews might cause changes in behaviour rates throughout the set, they are confounded and need to be empirically separated.

For the following experiment to assess whether external delays do cause a compensatory increase in subjects' pace, consistent with the proposal that they are escape motivated, the no-delay and noncontingent-delay conditions needed to be standardised in terms of any other features that delay completion of the materials, most notably review size. Review size has not previously been equated across conditions, however; No Delay has had relatively large reviews, whereas Noncontingent Delay has had smaller reviews. As a result, any motivation to escape would not necessarily reflect the influence of delays alone because reviews were not standardised across conditions, and this may have been a confound. The present experiment overcame this by removing reviews entirely. In this way, the present experiment was a systematic replication of Experiment 3 in which the influence of delay on pacing was isolated.

To determine the behaviours that are influenced by delays, the program used in the present experiment recorded each temporal component of a frame; that is, time required to read an item and respond, correct a response, and continue to the next frame (the latter two components were part of a combined measure in Experiment 3). This enabled the assessment of not only whether Noncontingent Delay led to a faster workrate, but also two alternative explanations of why this might have occurred. If, under imposed delays, time is reduced at only a select point (e.g., read and respond), then this suggests that a mechanism such as increased discriminative ability is responsible for time saving. If, however, each of these frame elements were reduced in Noncontingent Delay, then this would support the notion that some generalised motivational influence was operating to reduce session time (i.e., escape motivation).
Method

Subjects

Three undergraduate students were paid volunteers in this experiment: H44 was a 21-year-old, male, commerce and computing student; H45 was a 19-year-old, female, computing student; and H46 was a 21-year-old, female, biology student (see Appendix 2-2 for subject details). Payment contingencies were identical to those used in Experiment 3: Each subject received $5 for each session they completed plus a bonus payment of $50 if they did not miss any experimental sessions. If subjects missed one or more sessions, they did not receive the bonus and, additionally, lost $15 for each session for which they were absent. One subject (H44) had no experience with the subject matter, but the other two subjects (H45 and H46) had some knowledge of reflex behaviour (covered during the first half of baseline sessions) due to previous undergraduate studies. All subjects were naive with respect to the aims and rationale of the present study.

Apparatus

Design. Each subject received the following conditions: (a) No delay between items (No Delay), and (b) A 10-s delay after each item (Noncontingent Delay). These conditions were implemented in an identical fashion to the previous experiment except that there were no reviews. This meant that after the last question had been presented, answered, the correct answer had been shown, and the response had been corrected (and the delay had been presented in Noncontingent Delay), the program terminated. Questions answered incorrectly on the first pass were not re-presented.

Learning sets were presented over 20 days, two sessions per day, in the same fashion as Experiment 3: In the first five days Sessions 1 to 10 were presented with No Delay; in the next 10 days Sessions 11 to 30 were presented with either No Delay or Noncontingent Delay; and in the final five days, Sessions 31 to 40 were presented with No Delay. In the alternating conditions phase, conditions were assigned to sessions using a random sequence with the constraint that after two sessions both conditions had been presented. That is, the first session for any day had a condition randomly assigned, but the second session for that day was always assigned the other condition (see Appendix 4-2). This constraint was not employed in Experiment 3, in which there were two days when only one of the conditions was presented. The constraint of presenting both conditions on each day increases the likelihood that if background variables are influencing the results of one day, then any effects will be distributed across both conditions.
Retention. The pretest, posttest, and follow-up test used in the present experiment differed from those used in Experiment 3. For the present experiment, a 60-item test was developed. These items were randomly sampled from the test-item pool (see General Method) with the only restriction being that there were 15 items assessing the materials presented in each of the four conditions (Baseline, No Delay, Noncontingent Delay, Reversal). The full version of this test was employed as the posttest and follow-up test and a shortened version (7, 8, 8, and 7 items in the Baseline, No Delay, Noncontingent Delay, and Reversal conditions, respectively) was used as a pretest. The shorter pretest was considered sufficient because subjects, understandably, achieved very low scores.

Procedure. Each subject read the experimental instructions (see Appendix 2-4), completed the demonstration program, signed the consent form, completed the learning sets, tests, and debriefing as described in Experiment 3. The demonstration program had been changed so as to exclude the review feature.

Results

Accuracy

Figure 4.9 shows that all three subjects had high and reasonably stable levels of accuracy throughout the experiment. Differences in accuracy levels for No Delay (unfilled circles) and Noncontingent Delay (filled circles) were visually apparent for two subjects (H44, H46), but not the third subject (H45). For H44, data from the alternating conditions phase show that accuracy appears to descend across early sessions of No Delay, but recovers in later sessions. Accuracy is typically high and stable for Noncontingent Delay, and greater than that under No Delay, but sessions toward the end of the phase show that accuracy appears to decrease under this condition. For the second subject (H45), accuracy was high and relatively stable in all phases including both conditions of the alternating conditions phase. Visual inference suggests that the experimental variable has had no effect. For the remaining subject (H46), accuracy during baseline is variable but appears to decrease across sessions. This lower level of accuracy in No Delay is continued in the alternating conditions and reversal phases. Accuracy in Noncontingent Delay is typically higher, and visual inference suggests that Noncontingent Delay typically leads to better accuracy than No Delay.
Figure 4-9. For each subject, the percentage of frames correct per session, for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison.
Table 4-2
Proportion of Sessions in Which Noncontingent Delay had a Greater Proportion of Correct Responses (% Correct) than No Delay

<table>
<thead>
<tr>
<th>Subject</th>
<th>Self-scored % Correct</th>
<th>Rescored % Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>H44</td>
<td>6/10</td>
<td>6/9</td>
</tr>
<tr>
<td>H45</td>
<td>6/10</td>
<td>7/10</td>
</tr>
<tr>
<td>H46</td>
<td>8/10</td>
<td>8/10</td>
</tr>
<tr>
<td>Total</td>
<td>20/30*</td>
<td>21/29*</td>
</tr>
</tbody>
</table>

*p < .05.

Table 4-2 presents two versions of binomial comparisons of accuracy (percentage of correct responses) data: the self-scored data were obtained using accuracy as scored by subjects, whereas the rescored data were obtained using accuracy as scored by the experimenter. Both versions are included because they provide slightly different outcome patterns. The self-scored data show that Noncontingent Delay produced greater accuracy than No Delay for 20 of the 30 comparisons in which the conditions had different results (p < .05, sign test); this pattern also was evident in each of the three subjects (6/10 for H44, 6/10 for H45, and 8/10 for H46). This outcome was not compelling, however, because the delay advantage was only obvious in one subject (H46). Furthermore, although self-scoring accuracy was greater than 97% for all subjects (see Appendix 2-13), the few inaccuracies that did occur were biased in favour of No Delay. To overcome scoring bias, subjects’ responses were rescored by the investigator, and the percentage of correct responses for each set was recalculated. Using these figures, the number of comparisons in which accuracy was greater under Noncontingent Delay than No Delay increased for two of the three subjects (from 6/10 to 6/9 for H44, and from 6/10 to 7/10 for H45), and led to an overall total of 21/29 (p < .05, sign test). The trend for accuracy to be greater under Noncontingent Delay compared to No Delay, both within and across subjects, is readily apparent, and the rescored data strongly support the conclusion that Noncontingent Delay results in a consistent increase in accuracy in all subjects.
In addition, Figure 4-10 shows that, across all frames in the alternating conditions phase, Noncontingent Delay produced 6%, 0.3%, and 8% more correct responses than No Delay for H44, H45, and H46, respectively. The mean and median advantages of Noncontingent Delay were 5% and 6%, respectively. With the exception of H45, these results are similar to those found in each of the previous experiments, and show that Noncontingent Delay does lead to substantial improvements in accuracy regardless of the presence of the review feature.

![Bar chart showing percentage of frames correct for H44, H45, and H46 across different conditions.]

**Figure 4-10.** Percentage of frames correct in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject.

**Accuracy Differences Across Experiments**

It is possible that delays may be less beneficial when reviews are absent. That is, the delay advantage may be smaller in the present experiment than in previous experiments. Although this cannot be determined without an experimental assessment of delays with and without reviews, the present results appear to provide some support for this notion. H45 is the first subject in four experiments not to show an improvement under Noncontingent Delay of at least a few percent; the previous smallest effect size was 3.6% for H22 (Experiment 1) and H27 (Experiment 2). In addition, the 6% to 8% benefit of delays in the other two subjects (H44 and H46) seems to be smaller than expected (previous
experiments indicated that delays improved accuracy by an average of 8% to 10%.

An argument against this possibility is that the improvement of H44 and H46 under Noncontingent Delay is comparable with effect sizes found in all previous experiments. The difference in accuracy between Noncontingent Delay and No Delay in Experiments 1 to 4 is presented in Table 4-3. In each of the first three experiments there were instances of accuracy benefits smaller than, and greater than, those occurring in H44 and H46; hence, the effect is not obviously attenuated in the present subjects. A plausible explanation of the failure of H45 to show a benefit of learning under delay is that she performed at a very high level in all conditions, and perhaps there was a ceiling effect. It is noteworthy that H22, H27, and H45 were clearly the best performers in the experiment in which they participated and it is possible that, under their study-specific circumstances, they were performing at their maximum, and that the opportunity for additional study did not provide further benefit (i.e., a ceiling effect). With the evidence available it is reasonable to conclude that noncontingent delays are beneficial regardless of the presence or absence of reviews.

Table 4-3

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Subject</th>
<th>Accuracy difference (NCD - ND)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H21</td>
<td>11.8</td>
</tr>
<tr>
<td></td>
<td>H22</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>H23</td>
<td>9.2</td>
</tr>
<tr>
<td></td>
<td>H24</td>
<td>7.5</td>
</tr>
<tr>
<td>2</td>
<td>H25</td>
<td>4.8</td>
</tr>
<tr>
<td></td>
<td>H26</td>
<td>12.1</td>
</tr>
<tr>
<td></td>
<td>H27</td>
<td>3.6</td>
</tr>
<tr>
<td></td>
<td>H28</td>
<td>13.7</td>
</tr>
<tr>
<td>3</td>
<td>H35</td>
<td>9.5</td>
</tr>
<tr>
<td></td>
<td>H36</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>H37</td>
<td>7.9</td>
</tr>
<tr>
<td>4</td>
<td>H44</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td>H45</td>
<td>0.3</td>
</tr>
<tr>
<td></td>
<td>H46</td>
<td>8.1</td>
</tr>
</tbody>
</table>
Efficiency (mean time per frame)

Mean time per frame was calculated by dividing total time required to complete all frames by the number of frames (i.e., only the first pass was employed in this experiment). Two factors decrease efficiency; spending time in reviews of incorrect responses, and externally-imposed delays. Compared with previous studies, subjects in the present experiment should show improved efficiency because they did not lose time repeating errors. Furthermore, Noncontingent Delay should be far less efficient than No Delay because it still has imposed delays, but No Delay no longer loses time repeating errors. Due to imposed delays, frames completed in Noncontingent Delay should be, on average, 10 s longer than those completed in No Delay. If the magnitude of the difference in mean time per frame is less than 10 s, it will be entirely a function of a quicker work rate under these conditions.

Figure 4-11 shows the mean time per frame for each session. For each subject, mean time per frame under No Delay was stable (i.e., did not increase or decrease) across each phase of the experiment. Mean time per frame did, however, show a moderate level of variability which is not unreasonable considering the variation in set sizes and difficulty that existed. Mean time per frame under Noncontingent Delay also showed a moderate level of variability in each subject, and was consistently of a higher level than No Delay. Visual inspection of the data strongly suggests that mean time per frame is longer under Noncontingent Delay. This inference is supported by binomial comparisons: Across all subjects, the proportion of sessions in which Noncontingent Delay had a greater mean time per frame than No Delay was 29/30, which is significant (p < .01, sign test). This pattern was consistent across all subjects (10/10 for H44; 10/10 for H45; and 9/10 for H46). Clearly, Noncontingent Delay is consistently less efficient than No Delay when incorrect items do not need to be repeated.

Binomial comparisons show what happened on a consistent basis: for frames in each matched data set, an imposed delay added more time to a frame than could be regained by working quickly. This does not mean that, overall, no time was regained. Figure 4-12 shows that for two subjects (H44 and H45) the difference between Noncontingent Delay and No Delay in mean time per frame is approximately 9 s, whereas for the remaining subject (H46) the difference is 6 s. That there was some reduction from 10 s suggests that subjects responded more quickly in Noncontingent Delay.
Figure 4-11. For each subject, the mean time per frame in each session for No Delay (ND; unfilled circles) and Noncontingent Delay (NCD; filled circles).
Figure 4-12. Mean time per frame in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject.

First-pass Time and Time Components

In Experiment 3, AQT (time taken to correct a response and move onto the next item) was imprecise because it encompassed two different behaviours (correcting an item, and continuing to the subsequent item). This was rectified in the present experiment by recording separate measures of each of those behaviours. The controlling program recorded the following components of each frame: (a) Question Response Time (time required to read the problem and emit a response; QRT), (b) Answer Score Time (time taken to score a response as correct or incorrect; AST), and (c) Score Question Time (time between scoring a response and moving onto the next item; SQT). The sum of these components is First-pass Time (time required to complete a frame during the first pass through a set, excluding programmed delays).

Table 4-4 shows that, across all subjects, Noncontingent Delay had a greater First-Pass Time than No Delay in only 8 of the 30 comparisons, and this result is significant ($p < .01$, sign test). This outcome indicates that, in terms of contact time with the materials, subjects typically spend less time working on frames in Noncontingent Delay than No Delay. This pattern of results occurred for each of the three subjects.
Table 4-4

Proportion of Sessions in Which Noncontingent Delay had a Greater
First-pass Time, Question Response Time, Answer Score Time, and Score
Question Time than No Delay

<table>
<thead>
<tr>
<th>Subject</th>
<th>First-pass Time</th>
<th>QRT</th>
<th>AST</th>
<th>SQT</th>
</tr>
</thead>
<tbody>
<tr>
<td>H44</td>
<td>3/10</td>
<td>3/10</td>
<td>4/10</td>
<td>5/10</td>
</tr>
<tr>
<td>H45</td>
<td>3/10</td>
<td>3/10</td>
<td>3/10</td>
<td>3/10</td>
</tr>
<tr>
<td>H46</td>
<td>2/10</td>
<td>1/10*</td>
<td>3/10</td>
<td>0/10**</td>
</tr>
<tr>
<td>Total</td>
<td>8/30**</td>
<td>7/30**</td>
<td>10/30*</td>
<td>8/30**</td>
</tr>
</tbody>
</table>

Note. Comparisons are based on mean values of First-pass Time, QRT, AST, and SQT obtained in each session.
Mean SQT scores for Noncontingent Delay have had 10 s subtracted to adjust for the enforced delay.
Sign tests were performed only on total scores.
*p < .05. **p < .01.

Data relating to time components show that delays result in faster responding at each point in a frame: Across all subjects, Noncontingent Delay consistently had a shorter QRT (p < .01, sign test), AST (p < .05, sign test), and SQT (p < .01, sign test) than did No Delay. With only one exception (SQT of H44), this pattern was evident for each subject. Imposed delays have a general hastening effect, they typically result in faster responding, correcting, and continuing; this supports the escape proposition.

It is important to clarify the manipulation of the 10 s imposed delay in the SQT calculations. In a no-delay condition, a subject typically takes 1 s or 2 s to respond to the prompt and move onto the next item (SQT). When a noncontingent delay was applied, however, SQT included both the “user time” plus the additional 10-s imposed delay. This “artificial” increase in SQT was subtracted before any comparisons across conditions were made. For example, 2 s SQT under No Delay would be compared with 1 s SQT under Noncontingent Delay (which is calculated by subtracting 10 s from the 11 s SQT recorded). Hence, although it was reported that SQT’s were reduced under Noncontingent Delay, it is difficult to gauge how meaningful this result is because the independent and dependent variables overlap. However, SQT as defined here, does follow the
same pattern (smaller under Noncontingent Delay) as the other components of First-pass Time, and is consistent with the escape motivation proposition.

Efficiency and timing data indicate that, in Noncontingent Delay, some time can be recovered by working more quickly at each point in a frame. This condition is still less efficient than No Delay, however, and the size of the difference between conditions is considerably larger in the present experiment than in previous experiments in which reviews were present. This indicates that time lost repeating items in No Delay was the main factor responsible for similar efficiency of the two conditions in previous studies. The following chapter presents experimental assessments of the effects of reviews on all outcome measures, including the time cost per frame of repeating items.

Satisfaction

Session-by-session ratings and binomial comparisons for satisfaction are not included in this analysis because they were not consistent with subjects’ general opinions expressed during debriefing. Instead, general opinions and mean satisfaction ratings across sessions are presented.

Debriefing reports showed that subjects in this experiment had similar opinions about the experimental conditions as did subjects in Experiment 3; they were satisfied with both conditions, but preferred No Delay because Noncontingent Delay unnecessarily delayed session completion.

Figure 4-13 shows that, across all sessions, each subject rated both conditions highly, and that H44 and H45 rated No Delay higher than Noncontingent Delay by 0.7 and 1.2, respectively, and that H46 rated Noncontingent Delay higher than No Delay by 0.4. It is interesting that H46 claimed to prefer No Delay on the grounds that delays were distracting and broke her concentration - she actually performed 8% better in Noncontingent Delay.

Although Noncontingent Delay consistently rates high in satisfaction, it is apparent from the comments of several subjects that 10 s is considered longer than necessary to study the material, and if the duration were reduced, delays may be more popular. Other comments by subjects were that delays are used only after incorrect responses. This suggests that response-contingent delays would lead to the same accuracy improvement as noncontingent delays. This cannot be reconciled with the findings of Experiment 1, however, in which accuracy was greater in Noncontingent Delay than Contingent Delay. Perhaps when delays are noncontingent, subjects use them more than they realise or, because subjects are delayed more, they have increased escape motivation and concentrate harder when access to materials is returned.
Figure 4-13. Mean satisfaction ratings in baseline (Base), each experimental condition (No Delay, ND; and Noncontingent Delay, NCD), and reversal (Base2), for each subject.

Retention

Figure 4-14. For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The posttest and follow-up test were identical and 60 items long, and the pretest used half these items.)
Figure 4-15. For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). A missing bar indicates a score of zero. (Baseline, Review, No Review, and reversal were represented by 15 items each in the posttest and follow-up test, and by 7, 8, 8, and 7 items, respectively, in the pretest.)
Figure 4-14 shows that subjects performed substantially better in the posttest than the pretest, indicating that entering behaviour was substantially increased by the learning sets. Also, follow-up scores were equal to or higher than those of the posttest, showing that gains were well maintained over one month. These results are slightly lower than those of Experiment 3, perhaps because the present subjects did not have the benefit of repeating incorrect items or completing intermittent review sets.

Figure 4-15 shows each subject’s test performance for each experimental phase. In the posttest and follow-up test, materials learned during baseline were generally recalled best and, for the alternating conditions phase, there was typically a better recall of material studied with delays than without delays: H44 recalled 20% more noncontingent-delay material than no-delay material in both the posttest and follow-up test; the same pattern of results occurred for H45, although the magnitude of the effect was 7%. Similarly, H46 also recalled more Noncontingent Delay material in the follow-up test, but in the posttest, No Delay fared better.

Experiment 3 provided evidence that materials studied with noncontingent delays were well maintained whereas materials studied without delays were not, and that this caused an increase in experimental effects over time. This pattern of results was not replicated in the present experiment because two subjects (H44 and H45) had similar retention for both no-delay and noncontingent-delay material in both immediate and delayed retention tests. Materials studied under No Delay are not necessarily more poorly maintained.

**Constraints on Experimental Effects**

To this point, no attempt has been made to integrate data from all subjects because design differences across experiments would make any conclusions tentative at best. Nevertheless, Experiment 4 was the final direct investigation of delay, and it seemed reasonable to combine suitable data from Experiments 1 to 4 to detect any interesting relationships, even if they are only speculative.

Previously it was shown that some subjects who were very accurate under No Delay showed small gains under Noncontingent Delay (i.e., H22, H27, and H45), and this supported the notion that there is a ceiling to the advantages of delay. It seemed reasonable to use the data from the 14 subjects in the first four experiments to check the following proposition: If external pacing increases study, then subjects who performed poorly when self paced (and, hence, had more room for improvement) might make more use of delays, and show a relatively large experimental effect. Conversely, subjects who performed well under self-
pacing conditions may benefit less from delays and show a smaller experimental effect.

**Figure 4-16.** For each subject in each experiment, the mean accuracy across the alternating conditions phase for Noncontingent Delay and No Delay.

Figure 4-16 shows, for each subject, mean accuracy for Noncontingent Delay and No Delay (Noncontingent Delay is plotted as though the bar is behind that of No Delay). The means show the average scores obtained in the alternating conditions phase. Three things are evident from this figure: (a) subjects who perform well under one condition tend to perform well under the other condition; (b) for each subject (including H45), Noncontingent Delay led to greater accuracy than No Delay; and (c) when accuracy under No Delay was relatively poor, the benefit of delays was relatively large; conversely, as accuracy under No Delay increased, the benefit of delays tended to decrease. That is, those subjects who did not have much to gain, did not; perhaps because a ceiling effect occurred. The ceiling proposition is intuitively reasonable, has some empirical support, and is a basis for further investigation.
Additional Findings

Accuracy rates were quite high for two subjects: H45 made approximately 12% error regardless of the condition, and H44 made approximately 12% error in the noncontingent-delay condition and 18% error in no-delay sessions. This is impressive considering that incorrect items did not have to be repeated (i.e., a review feature is not necessary to motivate subjects to be accurate in the first pass), that review sets (Sets 17, 29, and 41) were removed from the study material, and that subject payment was not contingent on accuracy. The remaining subject (H46), however, made considerably more errors: over 20% in Noncontingent Delay, and over 30% in No Delay.

In addition to being very accurate, H44 and H45 typically completed sets more quickly than the estimated times provided in the text. This is to be expected because these subjects did not complete reviews. The remaining subject (H46), however, took longer to complete sessions than is estimated in the text. This finding, in addition to her relatively poor accuracy, suggests that she was not of the same academic calibre as the other students (her Grade point average was 0.7).

With the exception of H46, the performance of subjects in Experiments 3 and 4 was consistent with that expected from well programmed materials, and this increases the external validity of the present findings. The performance increases relative to Experiments 1 and 2 are probably due to using shorter experimental sessions in the later experiments, and this indicates that they were positive changes.

Summary

Accuracy was typically high, and was greater in Noncontingent Delay than No Delay. This is the fourth consecutive experiment to show the delay advantage, and the first to demonstrate that the advantage is independent of the presence of reviews. Test accuracy also was typically better for materials learnt under Noncontingent Delay. No Delay was much more efficient than Noncontingent Delay, probably because the lack of reviews saved this condition more time than it saved Noncontingent Delay. Even so, in Noncontingent Delay, subjects saved some of the time lost in imposed delays by working faster, and the finding that each component of a frame was completed more quickly in the delay condition supports a generalised motivating factor. Subjects rated both conditions high in satisfaction, but considered delays to be annoying when their duration exceeded the time required to consolidate or correct responses.
Discussion

One aim of the present study was to replicate the finding of increased retention of materials studied under noncontingent postfeedback delays. This did occur. In addition to improving accuracy in learning sets, delays improved retention of the materials studied under this condition in both an immediate posttest and a one-month follow-up test. This replicates the findings of Experiment 3, and indicates that brief, imposed delays have both short- and longer-term advantages. As discussed in Experiment 3, however, it is possible that the retention benefits associated with Noncontingent Delay are due, in part or wholly, to the novelty of the condition. Although study opportunity accounts for the accuracy advantages and is a logical explanation for retention advantages, an empirical resolution is warranted.

A second aim of this study was to assess the effect of delays upon workrate in the absence of reviews. (Review size had previously been a potential confound). Subjects did save between 1 s to 4 s per frame by responding more quickly under Noncontingent Delay, and each response component that was timed (QRT, AST, SQT) was shorter under delay conditions, indicating that this manipulation does have a generalised motivational influence on subjects.

It is important to recognise that, in Experiments 2 and 3, the presence of a review may have promoted faster workrates in No Delay: Although review size was uncontrolled, it was typically larger in No Delay, and this may have increased escape motivation and reduced First-pass Time. Delays, however, were more influential. In those experiments and Experiment 4, there is still a clear pattern of faster responding in Noncontingent Delay: 9 of the 10 subjects whose workrate (First-pass Time) was recorded in Experiments 2, 3, and 4 showed the pattern of faster responding in Noncontingent Delay. For these 10 subjects, delay lead to faster workrates in 66 of the 97 comparisons of the delay and no-delay conditions.

Remaining issues

The first four experiments demonstrated the effectiveness of imposed delays in improving accuracy and retention. That subjects can do this while functioning at a quicker on-task rate is important from an educational point of view. Of course, in Noncontingent Delay, subjects have had the benefit of small amounts of study, but the evidence suggests that they have an unutilised potential that may be accessible through the thoughtful use of contingencies.

Experiment 4 was important because it showed that the performance benefits, previously attributed to external pacing, were due to delays alone, and
not determined by review size. In some ways, however, this experiment was only
the first step in resolving whether reviews influenced performance in previous
studies. To comprehensively determine the effects of reviews, the second step is
to experimentally compare performance under review and no-review conditions.
If they are not different, then the outcomes of Experiments 1, 2, and 3, can be
more confidently attributed to delay. The next two experiments compared
performance under review and no-review conditions. Delays were not employed.
CHAPTER 5: EFFECTS OF REVIEWS

One feature often considered essential to programmed instruction and its implementation on teaching machines is the requirement that each item be answered correctly at least once. Because it was costly to implement in mechanical devices, however, manufacturers of teaching machines and experimenters often omitted a review function (Holland & Porter, 1961). According to the manufacturers, good programs should produce no, or very few, errors and repeats should be minimal, and so its omission was rationalised as inconsequential.

The extent to which reviewing incorrectly answered items influences learning is unclear because research in this area is limited. One experiment, however, has assessed the effect of reviews on test accuracy (Holland & Porter, 1961). In that study, two groups of students used teaching machines to complete the psychology program by Holland and Skinner (which is being used in the experiments reported in this thesis), but only one of these groups used a review. The review group performed 4% to 13% better than the nonreview group on retention tests (the advantage increased with material complexity). The review group also had slightly longer mean times per frame, but had first pass times which were equivalent to the nonreview group. The last finding suggests that students did not adjust their workrate to compensate for time lost in review. However, in this experiment the program was part of the students’ coursework, so their final grade depended on their performance, and they may have emphasised attention to materials rather than speed. Furthermore, the average time to complete the first pass through a set was 12.3 min, and the average time spent in review was only 1.3 min per set; brief sets and reviews may not prompt any compensatory increase in workrate. These data indicate that, even on programs with low error rates, a review can have a positive influence on retention. For this reason, Holland and Porter described the common exclusion of the review feature in commercial machines as regrettable.

Although Holland and Porter demonstrated that the review feature improves test performance they did not report any accuracy data for the learning sets; this makes it difficult to predict how reviews might influence accuracy in the following experiment. One experiment that did provide information about accuracy (Anderson et al., 1971), employed a computer to present 98 frames about the diagnosis of myocardial infarction from electrocardiograms to 188 subjects in eight experimental groups. Two of the groups, 100%-KCR (Knowledge of Correct Response) and Review, were most like Holland and Porter’s no-review and review conditions. It was found that the groups did not differ significantly in the number of first-pass
errors made, the first-pass time per frame, or on a criterion test of multiple-choice and constructed-response questions. On the criterion test the review group actually performed approximately 10% worse. One notable finding in this research was that intended punishers (time out and repeating incorrect responses) did increase the percentage of correctly-spelled responses, perhaps because the controlling program regarded typographical errors as incorrect responses, so subjects avoided repeats by spelling correctly.

From these findings it is difficult to thoroughly understand the function of reviews. For example, if a review functions as a punisher of incorrect responses, or it negatively reinforces correct responses, then it should increase accuracy: Anderson et al. found no accuracy effects, and although Holland and Porter did not report this information, they found retention effects that suggest improvements in accuracy. Reviews do delay completion of materials (like imposed delays), so would be expected to reduce First-pass Time. Neither of the previous investigations found this, however. It is possible that reviews serve only an educative role, providing practice after completion of the first pass. This would explain why first-pass measures (accuracy, First-pass Time) have shown no differences between review and no-review groups, but measures that reflect the entire set (first pass plus review) such as mean time per frame, and retention measures, show differences.

The following experiments in this thesis (Experiments 5 and 6) compared the effects of review and no-review conditions using the same outcome measures employed in previous experiments (accuracy, First-pass Time, etc.). Experiments 5 and 6 are similar to that of Holland and Porter’s because the same program and experimental manipulation were used, and because test items were taken from the same item pool (although the selection procedure was different). An important difference, however, was the use of a single-subject design in the experiments in this thesis. Whether this might alter the findings is unclear.

In addition to providing information of general interest regarding the effects on performance of repeating items, the following investigations of reviews will provide information on two issues directly relevant to this thesis. The first refers to the rate at which subjects work. Because reviews extend session length, they may contribute to escape motivation. By isolating reviews from delays, their influence on First-pass Time and its component measures can be determined. The second issue refers to the time cost that arises from manipulations designed to improve performance. The results of Experiment 4 showed that delays have a time cost of 6 s to 9 s per frame. Experiment 5 will assess the time cost of reviews. In earlier experiments this cost could have been calculated statistically, but may have been
inaccurate because it was difficult to separate the relative contribution of delay and review to time taken. An empirical determination of the cost is preferable.

Based on the limited findings presented above and information derived from the previous experiments in this thesis, some tentative hypotheses were that the presence of a review would: (a) have no influence on accuracy, or reduce it minimally because subjects would attempt to avoid repeats and longer sessions; (b) have little influence on work rate (First-pass Time) because, unlike imposed delays, repeats are not salient, noncontingent obstructions to session completion; (c) decrease efficiency because time would be spent repeating items; (d) improve retention due to increased practice of the learning material; and (e) be considered less satisfactory than a condition without a review because session time would be extended.

Experiment 5

Method

Subjects

Three undergraduate students were paid volunteers in this experiment: H38 was a 19-year-old, female, biology major; H39 was a 29-year-old, male, biology major; and H40 was a 20-year-old, female, biochemistry major (see Appendix 2-2 for subject details). Payment contingencies were identical to those used in the previous two experiments: That is, subjects were paid $5 for each session completed, penalised $15 for each session they were absent, and received a bonus of $50 for not missing any sessions. Due to their academic background, each subject had some knowledge of reflex behaviour (covered during the first half of baseline sessions), but they were not familiar with the rest of the material presented in the program. All subjects were naive with respect to the aims and rationale of the experiment.

Apparatus

Design. Each subject received the following conditions: (a) repeating incorrectly-answered frames at the end of a set (Review), and (b) not having to repeat incorrect frames at the end of a set (No Review). The present experiment employed the same design used in Experiment 4: Sessions 1 to 10 were presented with Review so that the present baseline conditions were directly comparable to those of previous studies. Sessions 11 to 30 were presented with either No Review or Review and the assignment of sets to baseline and experimental conditions was identical to that used in Experiment 4 (see Appendix 5-1).
Postfeedback delays were not used in this experiment. Sessions 31 to 40 included a return to baseline conditions in which sets were presented with a review.

**Procedure.** The procedure closely resembled that of Experiments 3 and 4. On the first day, subjects read the experimental instructions (see Appendix 2-5), interacted with the demonstration program (which no longer presented delays but, instead, demonstrated review and no-review conditions), and completed the consent form and pretest. The following day, learning sets were commenced and continued for 20 days; the day after this, the posttest was completed. One month later, the follow-up test and debriefing were completed.

**Results**

Two subjects, H38 and H40, had brief illnesses during this experiment which resulted in their absence for between one and three days. The day prior to their absences, each subject did complete both sessions but their accuracy was noticeably affected, so the comparison for this day (i.e., one set under each condition) has been excluded from the following results.

**Accuracy (percentage of frames correct)**

Figure 5-1 shows all subjects had high accuracy in each phase (Base, AC, and Base2). There was a moderate amount of variability, most noticeable in H40, but this did not appear to be reliably associated with a particular phase or condition. Visual inference from the data in the experimental phase (AC) suggests that there is no consistent difference in accuracy between Review (unfilled circles) and No Review (filled circles) for any subject.

Self-scoring accuracy was greater than 98% for all subjects (see Appendix 2-13). Nevertheless, when subjects’ responses were corrected by the experimenter (as they were in all experiments), one subject (H39) showed a systematic scoring bias in favour of No Review. To adjust for this bias, the proportion of correct responses per set calculated by the investigator were used in the binomial comparisons reported. Across all subjects, accuracy was better under Review than No Review for 15 of 28 comparisons in which scores differed, which indicates that there is no statistically significant effect of either condition; this supports the visual inferences. Although this overall finding was consistent with the performance for two subjects (6/9 for H38, and 7/10 for H39), the remaining subject (H40) performed worse under Review than under No Review (2/9).
Figure 5-1. For each subject, the percentage of frames correct per session, for Review (REV; unfilled circles) and No Review (NR; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison. Sessions with missing data are marked “M”.
Figure 5-2 shows the accuracy data averaged over baseline, Review, No Review, and return to baseline conditions, and indicates that, for each subject, the overall number of correct responses produced in each condition was high, and accuracy on baseline materials was highest. In the alternating conditions phase, two subjects (H38 and H39) performed similarly in both Review and No Review and for only one subject (H40) is there an appreciable advantage for one condition: accuracy was 4% better under No Review. Because the review feature failed to improve accuracy in two subjects, and actually produced worse performance in another, it cannot be considered to function as a punisher of incorrect responses or a negative reinforcer of correct responses.

No Review was not expected to improve accuracy. This occurred for H40, however and, according to this subject's debriefing report, can be explained in the following way: Making mistakes was more aversive in No Review than Review because there was no opportunity to correct the error. Hence, to avoid appearing unintelligent, she tried harder to avoid mistakes during the first pass of No Review because she could not redeem herself by getting the correct answer on a second attempt.

Figure 5-2. Percentage of frames correct in baseline (Base), each experimental condition (Rview, REV; and No Review, NR), and reversal (Base2), for each subject.
Efficiency (mean time per frame)

Mean time per frame was calculated by dividing the total time required to complete all frames (i.e., first pass frames plus any repeated items for the review condition), by the number of frames in the set (i.e., first-pass frames). Because there were no programmed delays, only time required to repeat items should contribute to this measure of efficiency, and Review should show higher values of this variable. Figure 5-3 shows that, for each subject, mean time per frame was quite variable during each phase. Even so, in the alternating conditions phase, a difference between Review and No Review was apparent in two subjects: H39 and H40 were less efficient (increased mean time per frame) under Review.

Binomial comparisons support the visual inferences. Across all subjects, Review had a greater mean time per frame than No Review for 20 of the 28 sessions in which they differed, which is statistically significant (p < .05, sign test). This pattern of consistently spending more time on Review items did occur for each subject (5/9 for H38; 7/10 for H39; 8/9 for H40). Figure 5-4 shows that each subject had a longer mean time per frame in Review than No Review, and that this difference ranged from 2 s (H38) to 6 s (H40).

Accuracy and efficiency data indicate that use of a review did not assist subjects. Under Review, each subject took more time to complete the task, but accuracy was unaltered in two subjects (H38 and H39) and was reduced in the remaining subject (H40). To put the efficiency data into a broader perspective, an increase in time per frame of 2 s amounts to an increase of almost 1 hr across the 1700 items in the present study materials; an increase of 6 s per frame amounts to an additional 3 hrs being required to complete the materials. This is wasted time if it does not increase performance.
Figure 5.3. For each subject, the mean time per frame in each session for Review (REV; unfilled circles) and No Review (NR; filled circles). Sessions with missing data are marked "M".
Figure 5-4. Mean time per frame in baseline (Base), each experimental condition (Review, REV; and No Review, NR), and reversal (Base2), for each subject.

First-pass Time and Time Components

Table 5-1 shows that, across all subjects, there is no difference between Review and No Review in First-pass Time. Similarly, none of the subjects showed a trend of quicker responding under either condition. These results indicate that subjects work at a similar rate in each condition, and two conclusions can be drawn from this: (a) increased mean time per frame under Review (2 s to 6 s) represents the time cost of repeating items, and (b) Review does not increase escape motivation. Perhaps a faster workrate is only likely when “unfair” (i.e., not contingent on errors) enforced delays occur; particularly when the obstruction to session completion is made highly salient (e.g., by the Delay bar).

It also is evident from Table 5-1 that, across subjects, there were no differences between Review and No Review in QRT or SQT. AST, however, was consistently longer under Review (p < .05, sign test); two subjects (H38 and H40) spent more time scoring responses when incorrect items had to be repeated. There appear to be two possible reasons for this. It is possible that they were being cautious in the presence of an aversive stimulus. Under No Review scoring an item as incorrect was inconsequential but, under Review, the consequences were an extended session and repeated items. Alternatively, they may use this opportunity to attend closely to the question, response, and correct answer to reduce subsequent
errors in this session. Although there are consistent differences across conditions in AST, they are too small to influence First-pass Time.

Table 5-1

<table>
<thead>
<tr>
<th>Subject</th>
<th>First-pass Time</th>
<th>QRT</th>
<th>AST</th>
<th>SQT</th>
</tr>
</thead>
<tbody>
<tr>
<td>H38</td>
<td>4/9</td>
<td>4/9</td>
<td>7/9</td>
<td>5/8</td>
</tr>
<tr>
<td>H39</td>
<td>5/10</td>
<td>5/10</td>
<td>5/10</td>
<td>6/9</td>
</tr>
<tr>
<td>H40</td>
<td>5/9</td>
<td>5/9</td>
<td>7/9</td>
<td>6/9</td>
</tr>
<tr>
<td>Total</td>
<td>14/29</td>
<td>14/29</td>
<td>19/28*</td>
<td>17/26</td>
</tr>
</tbody>
</table>

Note. Comparisons are based on mean values of First-pass Time, QRT, AST, and SQT obtained in each session.
* p < .05.

Satisfaction

Figure 5-5 shows session-by-session ratings of satisfaction averaged for each phase and condition. Both Review and No Review were given moderate to high ratings and Review was rated .3, .9, and .1 higher than No Review for H38, H39, and H40, respectively. These results indicate that reviews were not considered aversive, or substantially less preferable than not having reviews.

During debriefing, subjects expressed two different opinions about the conditions: Two subjects (H38 and H40) reported that a review was not aversive (perhaps because they did not consider finishing a session quickly to be important), and that they preferred Review over No Review because it provided them with the opportunity to redeem themselves if they made errors. Apparently, mistakes were not embarrassing if they could be corrected later. On the other hand, the remaining subject (H39) did claim that a review was aversive (contrary to his satisfaction ratings, Figure 5-5) because it lengthened session time. He reported that he tried to reduce errors during the first pass to avoid repeats, but, in fact, he showed no accuracy differences across conditions.
Figure 5-5. Mean satisfaction ratings in baseline (Base), each experimental condition (Review, REV; and No Review, NR), and reversal (Base2), for each subject.

Additional Findings

Because delays were not employed in this experiment, the conditions closely resembled those of traditional programmed instruction. Figure 5-2 shows that the mean proportion of items to which subjects respond incorrectly during the alternating conditions phase is approximately 10% to 15%. This error rate is considerably lower than the values in Experiments 1 and 2 (see Figures 3-2 and 3-8), and, like the findings of Experiments 3 and 4, resemble the error rates expected from well-programmed materials. Although the present subjects were not working for course credit based on their performance, and they did not have to complete examinations on this material, they were highly accurate. In addition to accuracy, times required to complete sets were typically the same or shorter than estimated completion times provided in the text. The present subjects can be considered capable and motivated to perform well.

Retention

Figure 5-6 shows the percentage of correct responses for each subject on the pretest, posttest, and follow-up test. Although each subject had covered some of the program content in their degree studies, their low scores on the pretest indicate that the amount of relevant prior knowledge was minimal. All subjects performed
substantially better in the posttest than the pretest, and this improvement was largely maintained in the 1-month follow-up test.

![Bar chart showing percentage correct for pretest, posttest, and follow-up tests for subjects H38, H39, and H40.](chart.png)

**Figure 5-6.** For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The posttest and follow-up test were identical and 60 items long, and the pretest used half these items.)

Figure 5-7 shows for each subject, and each test, the percentage of correct responses for test items assessing material covered during baseline, alternating conditions, and reversal phases. Posttest data show that, for all subjects, material completed during baseline (Sessions 1-10) was recalled best. Of the material completed during the alternating conditions phase (Sessions 11-30), no-review material was retained better than review material by all three subjects, and the difference across conditions was approximately 10%. In the follow-up test, retention was again best for baseline materials, but there was no systematic advantage of either Review or No Review for materials completed in the experimental phase.
Figure 5-7. For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). A missing bar indicates a score of zero. (Baseline, Review, No Review, and reversal were represented by 15 items each in the posttest and follow-up test, and by 7, 8, 8, and 7 items, respectively, in the pretest.)
Greater posttest retention of materials learned under No Review is interesting because (a) it had disappeared (in 2 of the 3 subjects) by the time of the follow-up test, (b) for two subjects (H38 and H39) it occurred even though there were no corresponding differences across conditions in the learning sets, and (c) it contradicts expectations based on Holland and Porter's test data. Those researchers demonstrated that the review requirement (having subjects repeat incorrectly answered items until the correct response is made) leads to improved posttest retention of approximately 10%. There are at least two explanations for the contrary results in the present experiment:

1. The retention measure was biased. It is possible that, in the process of random selection of test items, one or more of the items selected to represent Review were significantly more difficult than corresponding items representing No Review. Furthermore, in Figure 5-7, small differences between conditions in the number of items correct are visually compelling, because performance is expressed as a percentage of the total number of items representing each condition, and, in these calculations, the denominator was not large (15 items for each of baseline, Review, No Review, and reversal). This bias explanation, however, does not explain why the advantage for No Review was not also present in the follow-up test.

2. Holland and Porter used a groups design, so each subject experienced only one condition. In the present experiment, however, each subject experienced both conditions. Because the experimental condition (No Review) was novel, subjects may have had increased attention to, and therefore a better recall of, those materials. Nevertheless, novelty may not have a durable influence (especially compared to question difficulty) and this could explain why experimental differences had vanished by the follow-up test. However, if increased attention to No Review did occur, it is unclear why accuracy was not increased by this condition.

Summary

Review did not improve accuracy in any subject. One subject actually had better accuracy under No Review because, according to her debriefing report, she made an increased effort when there was no second chance to correct mistakes. Review was consistently less efficient than No Review because items that were answered incorrectly had to be repeated. However, the rate at which subjects worked through items was equal in both conditions, indicating that reviews do not contribute to escape motivation. Posttest data were unusual because material studied under No Review was recalled better than material studied under Review. This effect had vanished in two subjects at follow-up. It is unclear whether accuracy
differences in the posttest are due to differences in item difficulty, the experimental condition per se, or its novelty.

**Discussion**

Review did not improve accuracy: It had no influence on accuracy for two subjects, and decreased it for the other subject. Anderson et al. (1971) also found that reviews do not influence accuracy. If test performance is based on learning, it is difficult to reconcile these findings (that a review does not increase accuracy) with Holland and Porter's finding that a review does improve retention. Possibly Anderson et al. (1971) undermined the potential for accuracy differences across conditions by using a computer program that was not tolerant of typographical errors. Their subjects may have been generally more attentive and careful not to err in both conditions. It also is possible that the design used in the present experiment restricted accuracy differences, and this will be discussed further below.

The current retention findings are contrary to those of Holland and Porter (1961), and are confusing. There is a consistent advantage of No Review in the posttest. This advantage makes sense for H40 because she also was more accurate under this condition, but there is no clear reason why the other two subjects should show a posttest improvement for materials studied under No Review. Possibly this is evidence for a previous argument that a novel condition (i.e., the one introduced after baseline) will attract more attention and lead to greater retention. If this were true, however, increased accuracy and follow-up performance in this condition also would be expected. Perhaps the posttest performances of H38 and H39 are neither substantial nor replicable, because the effect wore off by the follow-up test.

Timing data are more consistent across studies. As predicted, Review and No Review did not differ in First-pass Time in the present experiment, and similar measures in the previous studies also showed no experimental effect (Anderson et al., 1971; Holland & Porter, 1961). Although one subject (H39) considered repeats aversive and tried to avoid them, he showed no accuracy difference across conditions.

Also as predicted, Review decreased efficiency. The direction and magnitude of the efficiency effect is similar to that found by Holland and Porter (1961). In the present experiment, subjects required an additional 2 s to 6 s per frame to complete each frame in Review. Across many subjects, Holland and Porter found that a review increases task time by 1.3 min per 29-item set, which is the equivalent of 2.7 s per frame.
Implications for Delay

In Experiments 1, 2, and 3, it was apparent that two factors were responsible for efficiency differences between No Delay and Noncontingent Delay. Subjects lost time in No Delay because they made more errors and had to repeat these items in a review, whereas they probably gained time in Noncontingent Delay by working more quickly. The next two studies confirmed this: Experiment 4 established that subjects do work more quickly (1 s to 4 s per frame) under delay conditions, and the present experiment established that Reviews do entail a time loss of 2 s to 6 s per frame. Although imposed delays seemingly add 10 s to every frame across a set, up to 4 s per frame can be offset by the quicker workrate that often accompanies it; furthermore, up to 6 s per frame overall can be lost due to repeating items under self-pacing conditions. Hence, imposed delays are often not inefficient.

Using the available information, two conclusions can be drawn regarding the function of reviews in the current paradigm: (a) A review does not appear to function as a punisher of errors or a negative reinforcer for correct responses because it did not lead to increases in accuracy. Furthermore, two of the three subjects commented that reviews were not aversive; and (b) a review does not increase escape motivation because First-pass Time was equivalent in each condition. Hence, the improved accuracy and reduced response times (time components) previously found under Noncontingent Delay can confidently be attributed to imposed delays and not to reviews.

It is surprising that Holland and Porter found a strong retention advantage for Review, but the present study did not, when each experiment used the same materials and experimental conditions. The discrepancy may be accounted for by methodological differences between the studies. For example, because subjects in the present experiment had completed a baseline phase under Review before No Review was introduced, they may have been trained to respond with high accuracy, and then been unlikely to alter this behaviour when consequences for making errors became less aversive. It is important to ensure that the present findings accurately reflect the effect of the experimental variables, rather than some feature of the design. To ensure that this occurs, the next experiment provides a reassessment of the review feature using an altered design.
Experiment 6

Two methodological features of Experiment 5 may have reduced the likelihood of obtaining the expected effects of Review:

1. Holland and Porter showed that differences in test performance between Review and No Review decrease with easier items. Therefore, because more recent revisions of the program have lower error rates they may be less likely to demonstrate an experimental effect. The version used in the experiments of this thesis is more recent than that used by Holland and Porter, and this may explain some of the current lack of effects. Nevertheless, according to the first author (J. G. Holland, personal communication, April 28, 1992), the two program versions are still highly similar, which suggests that Experiment 5 should have produced results more closely resembling those of Holland and Porter.

2. The major difference between the Holland and Porter experiment and Experiment 5 was the current use of a single-subject design. This single-subject design necessitated a sequential presentation of conditions. To maintain a baseline comparable with previous studies, subjects completed the first phase with a review present, and the experimental condition was defined as the absence of a review. If, however, a review limits first-pass errors, then including it in baseline may have trained subjects to emit few errors. In later sets, when the review was removed, there was no accompanying changes in contingencies which would promote a change in the response patterns established during baseline. That is, subjects were trained to perform well in baseline and there was no advantage to increasing errors when No Review was introduced. Furthermore, the tendency to maintain baseline response behaviours (few errors) is fortified by social factors: All subjects in Experiment 5 considered it more important to get items correct than to complete a session quickly. This factor also would make it unlikely that in later sets students would perform worse than their baseline standards. Hence, the purpose of the present experiment was to determine whether order of presentation of conditions was relevant to the effects of Review on behaviour.

The following experiment was a systematic replication of Experiment 5 in which the order of conditions was reversed: Review was introduced in alternate sessions following a baseline of No Review. There are several reasons why it was prudent to reassess the function of Review: (a) Reviewing incorrect items may improve accuracy when introduced after a baseline of No Review because it introduces consequences for errors (e.g., repeating items, increased session length) that subjects have not previously experienced; (b) Binomial comparisons showed that reviewing incorrect items appeared to hinder H40 but not the other subjects,
and this indicates that Review can influence responding, but that this influence is fragile, may differ across individuals, and may appear under a different procedure; (c) If the functions presently attributed to Review were due to the experimental design and not the variable per se, then subsequent conclusions regarding the function of delays and reviews may be misleading; and (d) Understanding how changes in methodology influence subject responding is relevant to future use of single-subject experimental designs.

Method

Subjects

Three undergraduate students were paid volunteers in this experiment: H41 was a 21-year-old, female, literature major; H42 was a 25-year-old, male, architecture major, and H43 was a 20-year-old, female, biology major (see Appendix 2-2 for subject details). Payment contingencies were identical to those used in Experiments 3, 4, and 5. Due to her academic background, H43 had some knowledge of reflex behaviour (covered during the first half of baseline sessions). All subjects were naive with respect to the aims and rationale of the present experiment.

Apparatus

Design. Each subject received the following conditions: (a) repeating incorrectly-answered frames at the end of a set (Review), and (b) not having to repeat incorrectly-answered frames at the end of a set (No Review). Assignment of conditions to sets was the exact opposite of that used in Experiment 5 (see Appendix 5-1) so that Sessions 1 to 10 were presented with No Review, Sessions 11 to 30 were presented with either No Review or Review (in the opposite order to Experiment 5), and Sessions 31 to 40 were a return to baseline of No Review.

Procedure. An identical procedure to Experiment 5 was used.

Results

Two conventions are employed in the following section. First, graphs portraying session-by-session results use the convention established in previous experiments that the introduced variable is denoted by filled circles. Hence, in Experiment 5, No Review had filled circles, but, in the present experiment, Review has filled circles. Second, to make results from Experiments 5 and 6 directly comparable, both Review experiments present binomial comparisons which show the number of times Review led to a greater value than No Review.

Accuracy (percentage of frames correct)

Figure 5-8 shows for each subject, and each set, the percentage of frames answered correctly during the first pass through materials. Although each subject's
overall response pattern was different, one consistent finding across subjects was that there was no visually apparent differences in accuracy between Review (filled circles) and No Review (unfilled circles) in the alternating conditions phase. In addition, H41 initially responded with a moderate to high level of accuracy in the baseline phase, but this was quite variable. Accuracy decreased over time for that subject, to stabilise at a low to moderate level during the reversal phase. H42 maintained a moderate to high level of accuracy throughout the experiment, although performance also was quite variable for this subject. In contrast to the other subjects, H43 maintained high and stable levels of accuracy across all phases.

Table 5-2 shows that, across all subjects, Review produced greater accuracy than No Review for 19 of the 30 comparisons, which indicates that there was no statistically significant superiority for either condition, and this supports the visual inferences. There is, however, a trend towards more correct responses under Review for each subject, and, when errors in self-scoring were corrected and the results reanalysed, this trend became significant.

Although self-scoring accuracy was greater than 98% for two subjects (H41 and H43), and greater than 95% for the remaining subject (see Appendix 2-13), the errors that did occur influenced binomial comparisons. When error rates per session were corrected, and binomial comparisons were recalculated, the number of sessions in which Review had a greater number of correct responses than No Review increased from 6/10 to 8/10 for H41, and from 6/10 to 7/10 for H43 (see Table 5-2). Rescored data show that, across all subjects, accuracy was greater in Review than No Review for 22 of the 30 sessions in which the scores differed, which is significant ($p < .01$, sign test). This trend also was apparent in each subject and supports the notion that accuracy does increase when Reviews are employed.

The results of the binomial comparisons must be interpreted with caution. Statistically, they are of borderline significance, and significance actually depends upon the subject's, versus the investigator's, interpretation of whether some responses could be considered equivalent to the correct answer provided in the text. Hence, they do not provide convincing evidence that reviews influence accuracy. In the General Method it was declared that statistical tests were being employed to support visual inferences which were the primary means of interpreting experimental effects. By these criteria, there is no convincing advantage for Review.
Figure 5-8. For each subject, the percentage of frames correct per session, for No Review (NR; unfilled circles) and Review (REV; filled circles). During AC, conditions have been plotted in pairs to facilitate comparison.
Table 5-2

Proportion of Sessions in Which Review had a Greater
Proportion of Correct Responses (% Correct) than No Review

<table>
<thead>
<tr>
<th>Subject</th>
<th>Self-scored % Correct</th>
<th>Rescored % Correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>H41</td>
<td>6/10</td>
<td>8/10</td>
</tr>
<tr>
<td>H42</td>
<td>7/10</td>
<td>7/10</td>
</tr>
<tr>
<td>H43</td>
<td>6/10</td>
<td>7/10</td>
</tr>
<tr>
<td>Total</td>
<td>19/30</td>
<td>22/30*</td>
</tr>
</tbody>
</table>

* p < .01.

Figure 5-9. Percentage of frames correct in baseline (Base), each experimental condition (No Review, NR; and Review, REV), and reversal (Base2), for each subject.
Although the session-by-session data show no obvious advantage for either condition, when this information is collapsed to provide a mean accuracy score for each phase, an advantage for Review is more apparent. Figure 5-9 shows that, across all frames presented in the alternating conditions phase, subjects H41, H42, and H43, performed 3.2%, 8.5%, and 0.7% better, respectively, under Review than No Review. The consistent and low to moderate improvement due to Review supports the results of the binomial comparisons that there is some positive effect of reviews. It is also different from the outcome of Experiment 5 in which Review failed to increase accuracy in any subject.

**Efficiency (mean time per frame)**

Mean time per frame was calculated by dividing total time required to complete all frames, including repeated items for Review, by the number of frames in the set. Review should show higher values. Figure 5-10 shows that, for two subjects (H41 and H42), mean time per frame under No Review was of a moderate level and variability throughout the experiment, whereas, mean time per frame under Review was more variable and of a consistently higher level. Hence, visual inference suggests that Review was less efficient than No Review for these two subjects. Mean time per frame for the remaining subject (H43) was of a moderate level and variability in each phase and in each condition, and there was no visually apparent difference between conditions.

Across all subjects, Review had a greater mean time per frame than No Review for 22 of the 30 sessions in which they differed, which is statistically significant (p < .01, sign test). As shown in Figure 5-10 this pattern of spending more time on Review items occurred for two subjects (9/10, H41; 9/10, H42), and there was no difference between conditions for the remaining subject (4/10, H43). Figure 5-11 shows, for each subject, the mean time to complete a frame in each condition: Review frames took from 2 s to 13 s longer to complete than No Review frames. For all subjects, Review was less efficient than No Review.

In Experiment 5 it was concluded that the use of a review was inefficient because it increased time on task without any corresponding increase in accuracy. In the present experiment there is some evidence of a different relationship between accuracy and efficiency: For each subject, Review again resulted in an increased time on task overall, but also led to more correct responses. Furthermore, larger accuracy differences between conditions corresponded to larger efficiency differences (see Figures 5-9 and 5-11).
Figure 5-10. For each subject, the mean time per frame in each session for No Review (NR; unfilled circles) and Review (REV; filled circles).
Although it makes sense that more repeats take more time, a similar correspondence between accuracy and efficiency differences did not occur in Experiment 5. The explanation for this lies in the fact that efficiency is not determined solely by the number of errors made, but also depends on workrate. In the present experiment, subjects had idiosyncratic workrates: H41 consistently completed frames most quickly, H42 was slowest, and H43 had an intermediate pace. Hence, a reasonable explanation of the consistent relationship between the magnitude of accuracy and efficiency differences is as follows: H43 had few repeats and completed them quickly, so she extended mean time per frame in Review by a small amount; H42 had to repeat many items, took a relatively long time to complete each item, and, hence, had a large mean time per frame for Review; Finally, H41 had the greatest number of repeats but completed them quickly, so mean time per frame increased by only a moderate amount. In Experiment 5, different combinations of workrate and number of repeats meant that, across subjects, accuracy differences between conditions will not necessarily correspond to efficiency differences between conditions.

Because Reviews necessarily increase time on task, they will always be less efficient than No Review (unless subjects make no errors in Review or work extremely fast in this condition). Whether the extra investment of time required by Review is worthwhile depends on the value of the accuracy and retention gains.

Figure 5-11. Mean time per frame in baseline (Base), each experimental condition (No Review, NR; and Review, REV), and reversal (Base2), for each subject.
First-pass Time and Time Components

Table 5-3 shows that, across all subjects, there was no difference between Review and No Review in First-pass Time. This finding replicates previous assessments of this variable (Experiment 5; Anderson et al., 1971; Holland & Porter, 1961), and indicates that the presence of a review does not increase escape motivation. One subject (H43) showed a pattern of reduced First-pass Time under No Review. This was unusual, and cannot be readily explained. This subject did claim to prefer Review, so this may represent an attempt to minimize time in the least preferred condition.

Of the time components, the only noteworthy finding occurred for one subject (H42) for whom AST was longer under Review than No Review in 9/10 comparisons. That is, he consistently took longer to score responses in Review. The same outcome was found across all subjects in Experiment 5, and it was proposed that they may be exhibiting caution in the presence of an aversive stimuli, or using the opportunity to attend closely to the question, response, and correct answer to reduce subsequent errors (and time-consuming repeats). It is possible that the former explanation is relevant to H42. During debriefing he reported that he found reviews aversive. Perhaps to avoid repeats, he made numerous scoring errors by marking incorrect responses as correct. Because he was under pressure to adhere to the scoring criterion, it is possible that he spent extra scoring time in Review determining whether his response was an appropriate alternative.

Table 5-3
Proportion of Sessions in Which Review had a Greater First-pass Time, Question Response Time, Answer Score Time, and Score Question Time than No Review

<table>
<thead>
<tr>
<th>Subject</th>
<th>First-pass Time</th>
<th>QRT</th>
<th>AST</th>
<th>SQT</th>
</tr>
</thead>
<tbody>
<tr>
<td>H41</td>
<td>5/10</td>
<td>5/10</td>
<td>5/10</td>
<td>5/10</td>
</tr>
<tr>
<td>H42</td>
<td>6/10</td>
<td>6/10</td>
<td>9/10*</td>
<td>7/10</td>
</tr>
<tr>
<td>H43</td>
<td>2/10</td>
<td>5/10</td>
<td>4/10</td>
<td>4/9</td>
</tr>
<tr>
<td>Total</td>
<td>13/30</td>
<td>16/30</td>
<td>18/30</td>
<td>16/29</td>
</tr>
</tbody>
</table>

Note. Comparisons are based on mean values of First-pass Time, QRT, AST, and SQT obtained in each session.

* p < .05.
Satisfaction

Figure 5-12 shows mean satisfaction ratings across sets. Ratings were typically low (bottom half of the 9-point scale), and Review and No Review did not differ by more than 0.5 points for any subject. Although the generally low ratings could be interpreted as dissatisfaction with the instructional approach, all subjects maintained (during debriefing) that they considered this approach a very useful teaching method, and one that should be incorporated into their regular classes. In addition, they reported clear preferences: Two subjects (H41 and H43) preferred Review over No Review, and the remaining subject (H42) preferred No Review.

Subjects in the present experiment gave the same two reasons for condition preference as subjects in Experiment 5: H42 disliked Review because it extended the session. He regarded session completion more important than being accurate; his approach was exemplified by the comment “it was more important to get to number 30 than to get 30 correct”. He considered repeats to be a time-based punishment for errors and reported that he attempted to minimise first-pass mistakes in Review (accuracy data support this). H41 and H43 were more concerned with performing well than finishing quickly, they preferred Review because it provided an opportunity to redeem errors.

Figure 5-12 Mean satisfaction ratings in baseline (Base), each experimental condition (No Review, NR; and Review, REV), and reversal (Base2), for each subject.
Additional Findings

Two subjects, H41 and H42, made large numbers of errors (more than 30% and 20% of frames, respectively) during the learning sets (see Figure 5-9). These error rates are similar to those of Experiments 1 and 2, and are greater than those occurring in the subsequent experiments (3, 4, and 5), and this provides a basis for some speculation about causes of high error rates. First, after Experiments 1 and 2, it was proposed that 3-set sessions may result in fatigue and, hence, increased error rates. Although reducing session length probably had a positive influence on accuracy for most subsequent subjects, the large error rates of H41 and H42 during relatively brief 1-set sessions indicate that fatigue is an inadequate explanation of poor accuracy. Second, the high number of errors made by H41 and H42 cannot be attributed entirely to experimental conditions (such as noncontingent payment) because H43 maintained a low error rate, as did all subjects in Experiment 5 (where the same conditions were employed). A third explanation is that each individual’s overall accuracy is largely determined by their level of scholastic ability and motivation to avoid appearing unintelligent. Because the present concern is with the relative effects of the experimental conditions rather than subjects’ absolute level of performance, the poor performance of these subjects does not jeopardise the outcomes. Of interest is the possibility of a ceiling constraint similar to that discussed earlier with regard to delays: Review has only noticeably improved accuracy, when performance in No Review has been intermediate (Figures 5-2 and 5-9).

Retention

Figure 5-13 shows, for each subject, the percentage of correct responses for the pretest, posttest, and follow-up test. All subjects performed substantially better in the posttest than the pretest, and, although each subject exhibited a small (approximately 5%) decrement in retention in the one-month follow-up test, the improvement was largely maintained. Two subjects (H41 and H42) had unexpectedly low scores in the posttest and follow-up test. This may be due to an inadequate comprehension of material: the same subjects made large numbers of errors during learning sets. Another contributing factor is likely to be the absence of a performance-based reinforcement contingency. At the time of posttest, subjects had completed the majority of their experimental sessions and earned most of their money. There was little reason to make an effort in these sessions other than reinforcement from personal achievement. Subjects in previous experiments generally performed well under these circumstances, but perhaps H41 and H42 had less concern for their own performance. It was evident from the debriefing
reports of H42 that he invested minimal effort into his performance, and the same may have been true of H41.

![Bar Chart]

**Figure 5-13.** For each subject, overall percentage of correct responses for the pretest, posttest, and follow-up test. (The posttest and follow-up test were identical and 60 items long, and the pretest used half these items.)

Figure 5-14 shows for each subject, and each test, the percentage of correct responses for test items assessing the material covered during the baseline, alternating conditions, and reversal phases. For two subjects (H41 and H42), material completed under Review was recalled best in both the posttest and follow-up test. The magnitude of the Review advantage was inconsistent across subjects and test occasion, perhaps because percentage scores were low and were greatly influenced by small changes in absolute number of correct responses. These subjects also had higher accuracy in Review, so experimental differences in learning and retention measures correspond. The remaining subject, H43, recalled no-review material slightly better than review material (approximately 5%). This difference is small and also corresponds to the negligible difference across conditions in accuracy.
Figure 5-14. For each subject, and each test (pretest, posttest, and follow-up test), percentage of correct responses from each experimental phase (baseline, alternating conditions, and reversal). (Baseline, Review, No Review, and reversal were represented by 15 items each in the posttest and follow-up test, and by 7, 8, 8, and 7 items, respectively, in the pretest.)
Summary

All subjects were more accurate under Review than No Review: Although this was a statistically consistent effect across all subjects, the magnitude of the effect was substantial for one subject, small to moderate for another, and negligible for the remaining subject. When Review improved accuracy (i.e., H41 and H42), it also improved retention. Review did not alter workrate (First-pass Time), but did decrease efficiency (mean time per frame). One subject considered reviews aversive because it increased session time, but the remaining subjects preferred reviews because they provided a chance to correct first-pass mistakes.

Discussion

Because of their similarities, the two review experiments will be discussed together. These experiments were conducted to determine the extent to which having to repeat incorrect items influenced performance on outcome measures and, hence, whether earlier findings attributed to delays need modification.

Time Measures

Time measures were consistent across all subjects in each experiment.

Efficiency. Review was consistently less efficient (mean time per frame) than No Review. When inefficiency is not offset by increased accuracy or retention (Experiment 5), there is no advantage in employing reviews. In Experiment 6, however, review improved accuracy and retention for two subjects, but whether these gains are worth the time loss requires a value judgement. In their study of the review feature, Holland and Porter found that frame completion times were lengthened by approximately 2.7 s (averaged across many subjects), and this was considered a worthwhile investment to gain the test improvement their subjects showed. H41 and H42 had much larger increases in mean time per frame (9 s and 13 s, respectively), and it is unclear whether the accuracy and retention gains warrant this time cost.

Workrate. Both experiments found that Review does not influence workrate (First-pass Time) through the first pass. This has also been found in previous groups-based studies of Review (Anderson et al., 1971; Holland & Porter, 1961). Hence, there was no support for the notion that Review causes increased escape motivation.

The timing data have important implications for the delay studies: Reviews do not cause any compensatory increases in workrate, so previous increases in workrate under Noncontingent Delay can be attributed solely to the influence of imposed delays. Although reviews are like delays because they increase session
time, they may not affect pacing because, unlike delays, they are not salient (cf. Delay bar), noncontingent, and do not increase study opportunity.

**Accuracy**

It is difficult to draw general conclusions about the influence of a review on accuracy. All subjects in Experiment 5, and one subject in Experiment 6, showed no differences across conditions. However, two subjects in Experiment 6 did show low to moderate advantages of Review. Perhaps a greater number of subjects in that experiment would have provided a better indication of how reviews typically influence learners. Certainly, the influence is neither strong nor convincing, and has appeared only when Review is introduced after a baseline of No Review. This may occur because Review introduces new error contingencies: each incorrect response results in a longer session time and repetition of items. Under these new contingencies, reducing errors has advantages which were not present during baseline, so errors in Review decrease compared to No Review. In Experiment 5 however, reviews during baseline meant that subjects had been trained to perform well, and when No Review was introduced (and the error-based contingencies were removed), there was no advantage of increasing errors so accuracy did not change.

**Retention**

Only Experiment 6 provided any reliable retention findings across posttest and follow-up test: Two subjects showed better retention for the material studied under Review. Although these findings provide support for those of Holland and Porter (1961), the lack of advantage for Review materials in the other four subjects in Experiments 5 and 6 limits confidence in the ability of Review to reliably increase retention. The advantage of single-subject methodology is evident here in two ways: (a) It indicates that Review will have neither a strong nor reliable influence unless it is employed in a particular fashion (i.e., as the only condition in a groups design, or following a baseline of No Review); and (b) Even under circumstances when Review (or other manipulation) leads to a desired outcome in some subjects, it will not necessarily work for all subjects.

**Preferences and Accuracy**

An interesting finding in the Review experiments was the poor relationship between subjects’ opinion of a condition, their expressed strategy for working under it, and their actual behaviour under this condition. For example, in Experiment 5, both H38 and H40 expressed a preference for Review because it provided an opportunity to correct errors made during the first pass. Only H40, however, consistently made more errors in during the first pass in Review. Both
H39 and H42 considered reviews aversive because they extended the duration of the session. Only H42, however, minimised repeats by making fewer errors in this condition. Perhaps surprisingly, he did not work more quickly in this condition as other subjects have done to overcome time losses due to delay (e.g., Experiment 4).

A well-established finding in many diverse areas of psychological study is that the way people predict they will behave under certain circumstances, is not always the same as they actually do behave under those circumstances. This has been evidenced again in the present studies. To know precisely how people will behave under certain circumstances, it is their behaviour that must be measured rather than an inferred correlate, such as opinion.

Conclusions

Because review size had not been controlled in delay and no-delay conditions in Experiments 1, 2, and 3, it was a possible confounding factor. It was reasonable that the review requirement could influence any of the outcome measures. After two experimental evaluations of the influence of Review on the measures used in the delay studies, there are three important findings:

1. There is no evidence that Review influences workrate, so the previous finding that subjects work 1 s to 4 s per frame more quickly in Noncontingent Delay can be attributed solely to delays.

2. Reviews consistently decrease overall efficiency (mean time per frame increases by 2 s to 13 s per frame), so in this way, they are expensive to implement, especially if they produce no redeeming performance advantages.

3. The effect of Review on accuracy and retention is not clear. Review had a positive influence on 2 of the 6 subjects, but only under particular circumstances. This suggests that any influence of this variable in the first three delay experiments was, at best, weak, and there is no reason to believe that it had any differential influence across delay conditions. These findings, along with those of Experiment 4 (delays included and reviews excluded), indicate that the accuracy and retention benefits gained under Noncontingent Delay (Experiments 1, 2, and 3) can confidently be attributed to delays.
CHAPTER 6: GENERAL DISCUSSION

Effectiveness of Delays

Brief, externally-imposed delays consistently improved student accuracy. In the delay experiments, each of 14 subjects was more accurate in Noncontingent Delay than No Delay. The advantage due to delays was considerable: For three subjects, improvement was greater than 11%; for 10 subjects, improvement was between 3% and 9%; and for only one subject the advantage was negligible (< 1%). In addition, with only two exceptions (follow-up test of H35 and posttest of H46), Noncontingent Delay also improved test performance by 5% to 37%. It is important to remember that these consistent and often substantial improvements were a result of a relatively simple manipulation of the pacing structure. To place the findings in some perspective, Holland and Porter (1961) considered a retention advantage of between 4% and 13% sufficiently large to warrant the inclusion of a review feature in teaching machines even though it was quite expensive. External pacing, using delays, had comparable or larger gains in accuracy and retention and are inexpensive to program.

The performance advantages due to delays need to be weighed against any costs in consumer satisfaction and inefficiency. Imposed delays were not annoying. They were typically scored on the positive end of the 9-point satisfaction rating scale, and sometimes scored quite highly (e.g., Experiment 4). Compared to No Delay, Noncontingent Delay rated similarly in satisfaction in the early studies (Experiments 1 and 2), but slightly lower overall in the later studies (Experiments 3 and 4). Verbal reports also showed that delays were not disliked by subjects, but that briefer delays may have been preferable.

Whether imposing delays was inefficient (Mean Time per Frame data) depended heavily upon the presence of a review and the size of it. When subjects had to repeat many items in No Delays, the time required to do this offset the extra time required by imposed delays, so that there was little overall difference in efficiency across conditions (e.g., Experiments 1 and 2). Imposed delays were relatively inefficient when there were fewer repeats in No Delays (Experiments 3 & 4). In general, the performance advantages of employing delays seemed to outweigh any time cost.

The consistency and magnitude of the advantages of noncontingent Delays are more notable when certain aspects of the task are considered. In discrimination (e.g., Brackbill, 1964) and concept formation (e.g., Bourne, 1957) tasks, a postfeedback
delay assists learning, probably because subjects work on elements of the task like evaluating alternative solution hypotheses or consolidating stimulus-response relationships. In the current task (programmed instruction), it was possible that delays would fail to assist performance because small steps and ample prompting promote successful responding (i.e., a performance ceiling). Nevertheless, delays were advantageous, both when error rates were higher than is typically expected in programmed instruction (and a ceiling effect on accuracy was unlikely), as in Experiments 1 and 2, and when error rates were typical of well-programmed materials (and a ceiling effect on accuracy was possible), as in Experiments 3 and 4. Even with a good-quality program, for which correct responses are highly determined, the “normal” approach of self pacing can be improved by some external control of pace. Furthermore, performance improvements were demonstrated in tertiary-level students who were probably clever, in possession of good study techniques, and, in these ways, less likely to be assisted by small changes in external pacing than nontertiary students.

Activity During Delays

Accuracy improved when study material (i.e., question, response, and correct-answer feedback) accompanied delays (Noncontingent Delay) but not when delays occurred without the study material (Noncontingent Delay Blank Screen). It is likely that subjects used the enforced wait to study the on-screen information. The nature of the behaviour occurring during delay is difficult to determine precisely because it is not available for easy inspection. One explanation of the advantage of postfeedback delays is that subjects use the on-screen information to modify incorrect responses. However, some of the current findings indicate that more than error correction takes place. In Experiment 1, more frequent exposure to delay resulted in better accuracy: Noncontingent Delay resulted in 6% better accuracy than Contingent Delay which was 4% more accurate than No Delay; this relationship occurred for 3 of the 4 subjects (Figure 3-2). If postfeedback delays are used solely for error correction, then there is no reason why Noncontingent Delay and Contingent Delay should differ. The additional delays following correct responses appear to assist beyond error correction. Perhaps subjects use the delay time to reinforce (or strengthen or consolidate) the already correct stimulus-response relationship, or use this time to rehearse previous material because there is no need for error correction on the current trial, or, because
programmed instruction is progressive, subjects may relate the information in the current frame to that in previous frames. During debriefing, subjects in Experiment 1 could only suggest that they read the screen during the delay, and did not make any distinction between their activity in Contingent Delay and Noncontingent Delay. The first step necessary to establish experimentally the relationship between frequency of delay and accuracy is to systematically control the number of additional delays received beyond those contingent on error. This was not possible in Noncontingent Delay because each subject differed markedly in the number of errors they made, and, therefore, in the number of delays which followed errors (contingent) and those which followed correct responses (not contingent).

Subjects' comments regarding the function of delays included that they provided "a chance to think", and a chance for information to "sink in". These are very similar to previous descriptions of learner activity during (noncontingent) delays; for example, "think time" (Boehm et al., 1971), and "think about and digest the material" (Kline, 1992, p. 64). Although they sound logical, these descriptions are of inferred events that may mediate the observed relationship between the imposition of delays and resultant performance changes. They do not, however, assist in specifying and promoting behaviours that could improve performance in related tasks in the future.

Other activity that is more accessible to investigation includes direct monitoring of subject eye-motion (e.g., Doran & Holland, 1971) during imposed delays. Future research wishing to obtain a more elaborate topography of behaviour that occurs during enforced delays might use this approach.

The role of brief study periods may also be better understood by manipulating learners' preexisting skills, or the way in which material is presented during delays. The subjects in the present experiments were not trained to use the delays in any particular way. Perhaps because they had good study skills, they just needed to be paced appropriately to enable them to employ these skills. In other computer-based tasks (e.g., multiple-choice tests, problem solving) using normal subjects (i.e., not autistic, impulsive, or learning or behaviourally disabled), delays also improved performance without any pretraining in what to do during the delay. Much like subjects in the present experiments, learners diagnosed as impulsive respond too quickly when self paced, and external pacing using delays is thought to improve their performance by providing an opportunity to inspect the stimulus materials or evaluate alternative hypotheses before responding (e.g., Dyer et al., 1982). Research with impulsives also
found that although subjects were helped by delays, they were helped more by specific training in search skills to employ during the delays (perhaps because useful strategies were often not in the subjects repertoire, or, if they were, they had not occurred because competing impulsive responses were stronger). Normal subjects (like those in the present experiments) may also benefit from training in specific study techniques that could be employed in the delay period. Because the extent of assistance from delays will depend on the nature of activity during it, this is a reasonable focus for future investigations. For example, to what aspect of the screen display should students be directed for maximum benefit? Should this be a function of the correctness of their response, or is all study good study? Could observing behaviours be ensured or made more efficient by requiring a screen/mouse pointer to contact key frame components?

Motivation

Both anecdotal and experimental evidence indicated that subjects were escape motivated. They reported that when moving through the respond-receive feedback-continue sequence, they continued on to the next item as quickly as possible; often without properly assessing feedback. They did so because they expected to be correct (i.e., they did not need to stop and consider corrective feedback), and because they wanted to complete the materials as quickly as possible. The time components (QRT, AST, SQT) showed that when subjects were under delay conditions, they responded more quickly at every opportunity. This racing behaviour reduces performance under self-pacing conditions because subjects do not stop to consolidate, rehearse, or practice the material; perhaps most important is that they do not get the benefit of feedback. These problems are apparent in the comment of one subject: “You want to get to the end. It’s like getting essays back— you don’t reread them”. Although most subjects reported racing, some subjects (e.g., H45) maintained that they were concerned about performing well and not looking unintelligent, so took the time that was necessary to complete the items correctly.

Being escape motivated does not mean that students found the experimental sessions aversive. Verbal feedback, positive satisfaction ratings, and high levels of accuracy indicate that they enjoyed both delay and no-delay sessions and were motivated to do well. Escape motivation is a statement of relative reinforcement contingencies: Other activities outside the laboratory (e.g., friends in the cafeteria) were more reinforcing than studying. Competing with more reinforcing activities is a reality
that most instructors experience. Escape motivation is so common that "early dismissal for early mastery" has been recommended to instructional designers as a form of motivational feedback / reward (Leshin et al., 1992, p. 190).

**Minimal effort**

Rapid responding when pacing is unrestrained entails skipping important study. This type of behaviour adheres to the law of least effort (Anderson, 1970): Students will short circuit an instructional task when possible and, as a result, will fail to learn fully from a lesson; this is most likely to occur when they are tired or under pressure to work quickly. The present subjects were motivated to work quickly, and, under self-pacing conditions, they were not compelled to study the material, so they did not. (Reviews led to response-contingent increases in session length, but they were not enough to make subjects slow down and respond more carefully.) An extreme example of minimising effort was provided by Geller (1992). He described a system in which students in his lecture could use push-button switches to control lights on a feedback panel, and signal whether the lecture pace was too fast or too slow. He reported that various go-slow behaviours, even saying nothing for 30 s, were unsuccessful in producing a majority judgement of "too slow". These students also appear to be following the law of least effort: The less content they received, the less they had to learn for exams, and the easier it was to obtain credit.

Clark (1982a) related subjects’ attempts to minimise effort with their satisfaction: Students will enjoy a method that they perceive will bring “maximum achievement with less investment of time and work” (p. 97). This reasoning probably accounts for many of the present subjects’ reported preference for a self-paced condition. Because no-delay conditions appear to require less time investment than external pacing, but are probably seen as requiring a similar amount of work, they are preferable. This preference for self pacing is unlikely to be modified by lower achievement in this condition because the subjects were unaware that delays improved performance. (The issue of the subjects' awareness of the effects of delays is discussed further below.)

Direct evidence of minimising effort has been found previously in programmed instruction. When Doran and Holland (1971) monitored subjects’ eye movements as they completed programmed materials, they found that subjects initially fixate at or near the response blank and then scan the text for key words and phrases. Rather than read the item in full, they hunt for clues; supposedly to avoid reading more than is necessary to provide an answer, and to finish as quickly as possible.
Another reason learners are unlikely to slow down of their own accord is that they do not realise that this can be beneficial: Even after experience with Noncontingent Delay, subjects were, without exception, surprised to learn that they performed better with delays. An interesting issue concerns how much the computer medium contributes to this racing behaviour, and whether this was much less a concern when students self-paced on mechanical devices (e.g., the machines of Pressey and Skinner).

Solutions

How to deal with escape motivation and attempts to minimise effort are important concerns. External control of pacing (by imposing brief study delays) was effective in stopping learners from skipping an important study opportunity, but, in some ways, might be considered an artificial strategy. An alternative solution might be to ensure that programs are so skilfully constructed that they would guarantee studying automatically. That is, correct responding would be extremely unlikely without closely attending to the material, an effective study pace would be adopted without additional intervention (such as delays), and responding would be largely, or completely, error free. In some ways, however, this solution is unrealistic. Writing programs so that responses are highly determined takes a great deal of time and effort, and this is unlikely to be carried out by many program writers. Even when programmed instruction was popular, many programs were written that were of poor quality (Kemp & Holland, 1966). Given the history of quality problems in CBI, it is probable that many programs of dubious quality will continue to be produced. Even magnificent programs are unlikely to be as reinforcing as activities available outside the classroom or lab.

At first it might seem obvious that the way to compete with more reinforcing activities (in this case, nonclassroom activities) and minimal effort is to consequence high levels of performance with strong reinforcers (such as money). However, this may not necessarily be successful. For example, Moore and Smith (1964) had several groups of subjects complete the Holland and Skinner program under one of a variety of forms of feedback, including a group who received 1 penny per correct response. Although this group’s accuracy improved over a group without feedback, it was not as high as a group who were informed of what the correct response should be. In addition, their test performance was no different from any of the other groups (who did not receive payment). Response-contingent payment for performance on the Holland and Skinner program was also evaluated as part of the pilot work for this thesis. It was found that 15 c per correct response resulted in accuracy of a level equivalent to, but not higher
than, that of Experiments 3 and 4 where payment was contingent on attendance but not performance. Other investigations using different materials and payment contingencies also have failed to find significant effects of financial incentives on performance (e.g., Lintz & Brackbill, 1966; Sullivan, Baker, & Schultz, 1967). These findings indicate that even employing a strong response-contingent reinforcer such as money is not necessarily more effective at improving learning than noncontingent consequences. Furthermore, money is an unrealistic reinforcer for academic environments.

It has been more common in classrooms to use aversive contingencies such as performance hurdles or doomsday contingencies, to stop students procrastinating and promote study. Doomsday contingencies operate across sets to inform the student that they must study successfully and meet certain deadlines, or else! They do not however, assist the student at the time they are actually studying. Pacing control by imposing delays is a complementary instructional tool. Delays assist within an instructional frame by prompting students about when to study and for how long.

Subjects in the present experiments demonstrated behaviours typical of students: they minimised time spent and effort whenever possible. Self-pacing conditions enabled them to race through materials without taking effective study opportunities, so they did: External pacing within frames improved performance by enforcing these opportunities. The present experimental procedure was closely related to many realistic instructional situations because reinforcement was weakly contingent on effort expended studying, and more reinforcing activities competed with study. Until program designers consistently make very high quality programs, students learn to employ brief study periods at appropriate times, schools can promote studious behaviour over speed and inaccuracy, or afford and implement effective response-contingent reinforcement, it is important to investigate other ways of influencing behaviour—especially when it is informative about the learning process, and leads to low-cost, easy implementations such as the use of brief delays.

Compensatory Pacing

There are two alternative explanations of the increased workrate under Noncontingent Delay. First, delays promoted specific behaviour such as increased discriminative ability. Alternatively, delays had a generalised influence on motivation. These alternative explanations may not be independent, however, and a combination of both may be the best explanation for all of the workrate data.
In Experiment 4, each response behaviour that was timed (QRT, AST, SQT) was faster under delay conditions than no-delay conditions, indicating that this manipulation does have a generalised motivational influence on subjects. Because their pace is externally slowed at one point, subjects work faster whenever they are in control of the pace. The findings from Experiment 2, however, show that escape motivation is not a complete explanation of workrate differences across conditions. Here, Noncontingent Delay and Noncontingent Delay Blank Screen employed delays of equal duration (i.e., subjects were equally escape motivated in each condition), but only the former condition had shorter First-pass Times. Furthermore, Noncontingent Delay Blank Screen and No Delay were no different in First-pass Time. This shows that something more than escape motivation is necessary for subjects to work more quickly.

It can be speculated that even when there is a generalised escape motivation (due to delays being present), faster workrates are possible only when subjects also have an increased ability to make discriminated responses (i.e., when study materials accompany delays). In Experiment 4 subjects were not only faster to respond (the primary time measure contributing to First-pass Time) under delays, but they also were faster on relatively straightforward behaviours (scoring and continuing) because these behaviours were part of a chain that benefits from increased discriminative ability and the accompanying accuracy and speed of responding.

To properly determine whether scoring and continuation rates can increase without study opportunities, it is necessary to empirically separate study opportunity from delay per se. Experiment 2 provided this separation, but did not record the different components of First-pass Time. Alternatively, equal-sized delays could be interposed at different points in the learning sequence: after the question, after the response, and after the feedback. If the delay conditions all lead to faster responding than self pacing, and the amount of speed up is a function of how much study assistance is provided by the manipulation (e.g., postfeedback delays, prerresponse delays, and prefeedback delays lead to fastest, intermediate, and slowest workrates, respectively), then this would support the importance of study in increasing workrate, not just the presence of delay.

The Self-Pacing Tenet

When early researchers (e.g., Pressey, 1926, 1927; Skinner, 1954) advocated self pacing, their impetus was the common situation of entire classes of students
moving at the same rate. Compared with this approach, allowing students to control their own pace was an improvement: slower students were not left behind, and faster students were not restricted. Evidence suggests, however, that unrestricted generalising of the self-pacing approach to instructional settings outside the traditional classroom is ill-conceived. Not only is there evidence from a variety of tasks (e.g., concept identification, paired associates learning, problem solving, and programmed instruction) showing that external pacing can improve performance, but research on pacing in a computer medium (e.g., Boehm et al., 1971; Dwyer et al., 1985; Stokes et al., 1988) and the present studies also support the notion that a degree of external pacing can be beneficial. After two decades of CBI implementation, however, little has been done to advance upon the self-pacing approach; perhaps because good ideas are not rapidly abandoned, because psychological research often progresses slowly, or because external control is behaviouristic and therefore not in keeping with current, cognitively-oriented approaches to instruction.

Pacing is not an all-or-none issue--neither complete external pacing (by the programmer) nor self pacing (by the learner) is optimal--so prescriptions for pacing control require refinement. Complete external control does not cater for individual differences in pace, or variations in pace within one individual on different items, content, or sets, and this may decrease performance. Complete learner control leaves pace-related behaviours such as speed of response, scoring, and continuance, and timing of changes in pace under the control of competing contingencies (discussed previously) and aspects of the medium (e.g., speed of computer operation) that may lead to reduced performance.

The current experiments demonstrate one way of striking a balance between complete external pacing and complete self pacing by making a minor modification to what is essentially a self-pacing approach to learning. Learners were given control over almost all pacing decisions except for a brief, externally-imposed delay at a key instructional point (when feedback enables the learner to consolidate their existing understanding, or modify an incorrect response). The current combination of internal and external control of pace is an extension of Higginbotham-Wheat’s (1990) suggested compromise between internal and external control in instruction. She proposed that students be given control over noninstructional aspects of the program (such as text density) to provide affective benefit, but that the program control instructional aspects (such as practice) to provide learning support. Pacing is a single
aspect of instruction, but the present subjects had the affective benefit of control over most pacing decisions, but the performance benefit from noncontingent delays.

Future investigations of pacing (and other components of control) could be profitably spent in investigating clever ways of combining both the control requirements of the learner with the instructional knowhow of programmers. Although it is easier to follow all-encompassing tenets, a science of instructional design will benefit more from devising a collection of strategies that can satisfy the affective and instructional needs of learners, and be suited to particular instructional tasks. Several alternatives to complete self pacing in CBI have now been investigated. Dwyer et al. (1985) and Canelos et al. (1985) approximated a tailored reading time by using the number of lines of text to determine the stimulus exposure duration. Hativa et al. (1991) demonstrated that extending opportunity to respond beyond purported appropriate intervals can be advantageous. This approach was counter to the more typical one in CBI of reducing response time to avoid procrastination. Because the time required to read a question and respond is typically the most lengthy component of CBI, this component of the stimulus-response-feedback learning sequence may provide most scope for pace manipulations. Another way of tailoring exposure time of materials to subjects, and one that may promote efficiency, is to record subjects’ own read and respond times during baseline or pilot sessions, and use this information to devise reasonable upper limits for stimulus exposure duration. Then quicker, accurate responding could be shaped by reducing this limit over subsequent sessions. The calculations and implementation for individual subjects would not be difficult if computers were being used as instructional vehicles. These pacing alternatives, along with imposing delays are not complete educational solutions, but do contribute more strategies for how to effectively combine modern devices of information transmission with the psychology of instruction.

The self-pacing tenet and external pacing alternatives have, in this thesis, generally been referred to in the context of programmed instruction or CBI (due to the materials, use of computers as an instructional vehicle, ubiquity of computers in education, and problems with software prescriptions). But the bottom line from the experiments is not restricted to these areas; it is relevant to all forms of individualised instruction. Whenever students work quickly on a task and this appears to contribute to them making more errors than is reasonable, their overall pace needs to be coupled with
some form of external pacing that will promote attention to the important aspects of the task.

Duration of Delay

A variety of investigations have shown that imposed delays can have positive or negative consequences for performance, and that the optimal delay duration is task specific. Dyer et al. (1982) found that autistic children's performance on a discrimination task improved with delays of 3 s to 5 s, but declined with delays of 6 s. Boehm et al. (1971) found that a delay of 5 min improved problem solving performance, but 8 min was substantially less effective. Stokes et al. (1988) found that 30-s response lockouts improved multiple-choice test performance more than 60-s lockouts. Whereas overly brief delays do not allow enough time for task-related activity, overly lengthy delays may cause subjects to forget important information, or be distracted or irritated.

In the present experiments, 10-s delays worked well with the programmed materials, but it is possible that a longer delay (study period) would have caused greater improvements, or that a shorter delay may have provided the same benefit in less time. During pilot work for the current studies, delays of up to 40-s were employed and, although the studies were not directly comparable to the present ones, there was no evidence that delays longer than 10 s would be more effective: In addition, there was consensus among these subjects that delays longer than 10 s were annoying.

It is possible that a delay shorter than 10 s may have produced the same levels of accuracy and retention. Some subjects reported that a 5-s postfeedback delay was long enough for any necessary study, and that the remaining wait was unnecessary or annoying. In Experiments 1 to 4, Noncontingent Delay took between 0 s and 9 s per frame longer than No Delay. If a further 5 s per frame could be removed from the delay condition, then this would largely reduce, and in some cases remove, any efficiency differences between the two conditions. Smaller delays may also increase subjects' satisfaction with the procedures.

One early investigation of fixed versus self-pacing in programmed instruction (Kress, 1966) has not been mentioned previously because it did not employ imposed delays within trials but, instead, controlled the total frame time. Nevertheless, the attempt to empirically determine an ideal pace is relevant here. Kress intended to improve the accuracy and retention of low-ability students by making them spend more
time on frames, and improve the efficiency of high-ability students by reducing their frame times. Frame completion times were recorded for sixth- and twelfth-grade students working under self-pacing conditions, and these were used to calculate hypothetical optimal times for each frame in a program. Other subjects then completed the program according to the pace dictated by these times. Externally pacing high-ability students in this way did decrease their time on task, but there were sometimes performance costs. More important to the current experiments was the finding that extending frame time had no effect on low-ability students; they continued to respond rapidly and then waited until they could proceed to the next frame. The extra time was apparently not spent in task-related activity, but was wasted on task-irrelevant behaviours. The benefit of the current procedure is that the extra time is strategically located (following feedback) rather than generalised.

In summary, although delays are beneficial, longer delays are not necessarily better than shorter delays. Using the shortest possible delay to obtain performance increases is desirable both for task efficiency and learner satisfaction. The benefits of delays in the present programmed materials were, to a large extent, due to their brevity: they were not so long that they irritated learners or gave them an opportunity to be distracted, but they were of sufficient duration to enable some effective study. Because subjects were eager to finish the session, they attended to the screen and waited for the next question.

Reviews

A review requirement during learning sets has been shown to increase test performance by up to 13% (Holland & Porter, 1961). This indicated that reviews functioned as mild punishers of incorrect responses because they increased session length. The effects of reviews on performance were investigated in this thesis because review size may have confounded the findings attributed to delays, and because it was important to obtain a direct measure of the time cost of repeating incorrect items at the end of a set.

Outcomes from these investigations (Experiments 5 and 6) were not conclusive. There were often small and unreliable performance differences across review and no-review conditions which indicated that the review feature did not have strong, if any, control over subjects’ behaviour. In Experiment 5, two subjects had similar accuracy in review and no-review conditions, and one subject was more accurate under conditions
of no review. For each of the subjects, posttest performance was better for material that had been studied without reviews, but there was no consistent advantage for either condition in the follow-up test. It appeared that a review requirement had no reliable influence on these measures, but the results of Experiment 6 provided slightly more consistent outcomes. For two of the three subjects, reviews increased accuracy, posttest, and follow-up test performance; for the remaining subject there was little difference across conditions on any of these measures. When review increased accuracy, it also increased test performance.

It is unclear why reviews were influential in only one experiment, and whether these two subjects provided representative outcomes. Based on the outcomes of Experiments 5 and 6, it was argued that a review requirement may improve performance only when it is introduced after a baseline without reviews. If this is true, then the review condition may have benefited from a contrast effect. It may have little inherent controlling ability.

The most consistent finding in the review experiments was that the mean time per frame was longer under review than no-review conditions (this was due to repeats because First-pass Time did not differ across conditions). Reviews were time costly. One redeeming feature of reviews was that some subjects liked the opportunity to redeem mistakes. This is not necessarily a good feature, however, if it promotes inaccuracy on the first (or other) pass. Other subjects actually preferred not to have reviews because they were time costly.

There are two main implications of these findings. First, reviews were unlikely to have had a confounding influence in the delay experiments. Second, there is little reliable evidence from the current experiments that programmed instruction is improved by the review requirement. This has further implications for designers of instructional materials that incorporate remedial procedures. What form of remediation to use with which materials is not known because there has not been a lot of systematic research conducted to date (Clariana, 1990). Demonstrating mastery of every item by repeating those answered incorrectly at the end of the set, may not be an effective procedure. It is possible that learners in programmed instruction are unconcerned about repeats because they can remember the response from the first pass. They are assisted in this because feedback provides the correct answer regardless of the quality of their response. They are also aided in reproducing the correct answer later because items are sequentially related, and there is much overlap in the content covered from item to item. When
performance levels are high, this is not problematic, but otherwise alternative procedures may be better. One alternative to end-of-set reviews in computer-based instruction is the use of software that will accept constructed responses, but not provide the correct answer or move to the next item until a learner constructs the correct answer. When multiple-choice response formats are used, learners can randomly select options until they find the correct answer. By using constructed responses, learners will have to actively apply what they have learned to continue further through the materials.

It remains unclear why Experiments 5 and 6 did not produce unequivocal support for the review procedure such as Holland and Porter (1961) obtained. Perhaps the most likely explanation is the design used; by using a groups design, Holland and Porter avoided any contamination of one condition by the other condition (e.g., establishing high performance levels in baseline conditions that would not be reduced under changed conditions). Because more than one remedial procedure is unlikely to be employed in applied settings, reviews may warrant further investigation with groups designs and realistic materials.

Time-On-Task Notion

The relationship between performance and efficiency in learning is contentious. One of the first areas of programmed instruction to be experimentally evaluated was the effectiveness and efficiency of alternative response modes (e.g., overt and covert constructed responses, multiple choice, and reading frames with responses filled in). A popular expectation which arose from this research was that “any instructional method which forces the student to spend greater periods of time on task is likely to lead to higher achievement than methods requiring less student time” (Tobias, 1973, p. 202).

There is a range of evidence, however, that instructional methods requiring larger investments of time do not necessarily increase achievement. For example, in the present experiments brief postfeedback delays and end-of-set reviews both increased time on task, but only the former reliably increased performance. In an investigation of feedback procedures used with programmed instruction (Anderson et al., 1971), it was found that, compared to a group similar to the no-delay condition of the present experiments (100%KCR), imposing a 15-s timeout (with the frame in view) after incorrect responses did increase mean time per frame, but did not increase accuracy or test performance. To test the efficacy of a related review procedure, Kinzie et al. (1988)
had two groups study introductory material about solar energy (using CAI), and answer multiple-choice questions that were interposed with the content. In one group (program control) an incorrect answer meant that the subject had to repeat the relevant content section before attempting the question again. In the other group (learner control), an incorrect question could be reattempted immediately; repeating the content beforehand was optional. The learner-control group only reviewed material 35% of the time (cf. 100% for the program-control group), but achieved the highest posttest scores and took no longer to complete than the program-control group. Better performance was not obtained by increasing content review, and did not require increased task time. These outcomes demonstrate that not just any instructional method that increases time on task will result in better performance.

In a similar vein to Tobias' (1973) comments, Hunt and Mathis (1966) proposed that "the single most important variable in learning is the time spent in learning... the implications of this for education are obvious" (1966, p. 143). Hunt and Mathis' emphasis on time spent in learning can be interpreted as subtly, but crucially different from, Tobias' time on task. It implies that the added task time must be actively used in learning. This is important because the findings mentioned above indicate that increased time does not necessarily mean better learning, and that whether additional task time should be spent, how it is spent (e.g., rereading content, reviewing questions, daydreaming), and where it is spent (preroutine, prefeedback, postfeedback) are problems without obvious solutions.

A different relationship between time on task and performance was suggested more recently by Tudor and Bostow (1991) as a result of their experimental findings. They measured the performance of undergraduates on a posttest (constructed response items) and an applied-task (writing program frames) following participation in one of several groups where programmed materials were used to teach program writing skills. They found that as more components of programmed instruction (e.g., none, covert, or overt responses required) were included in a group's instruction, both time and performance increased. In light of this, they maintained that "time spent is not a cause of learning but is rather a byproduct" (p. 367), and emphasised the activity, rather than the investment of time, as the active ingredient in learning. This is not only logical, but it provides a more useful focus for future work on the nature of the subjects' activity during learning than an emphasis on increased task time.
Nevertheless, Tudor and Bostow's observation suggests a one-directional relationship between task activity and time taken (the former causes the latter, and not the opposite) which may be misleading. Time on task and activity have a relationship somewhat like the proverbial chicken and egg; it is difficult to pinpoint which initiates, or is responsible for, the other. Although doing more task-related activity is typically associated with increased time, a highly-motivated subject may take the same or less time. Conversely, although increasing time on task (e.g., keyboard lockouts, imposed delays) can provide an opportunity for further work that does not exist without these techniques, an unmotivated or distracted learner may do relatively little work in the time allotted. Whether different interventions are effective depends on their ability to focus learner's attention (i.e., their observing and echoic behaviours; Holland, 1960) to the important aspects of the learning material. Stokes et al. (1988) would probably not have improved student performance on test items if they had delayed students after responding, rather than prior to responding. Similarly, imposed delays often do not aid performance when they occur between a response and feedback. Employing reviews may not work if students pay less attention to questions they have already attempted. Because students learn by performing task-related behaviour and being reinforced for it, it would seem most efficient and beneficial to investigate procedures and contingencies that actively involve students and improve their performance; the relative efficiencies of the successful techniques is a secondary concern.

Experimental Issues

Experimental findings are valuable only to the extent that the methodology responsible for them is sound and the work is meticulously conducted. In the present experiments, the single-subject design typically provided clear and unambiguous outcomes. The experimental conditions demonstrated control over behaviour that was detected visually and statistically. Moreover, this control was possible with realistic materials over an extensive course of instruction. The computer was a very useful research tool that systematically and reliably presented the experimental contingencies to subjects, and recorded their responses and response times with precision. In short, a rigorous assessment of academic behaviour under quasi-realistic conditions was conducted successfully. Nevertheless, there are aspects of the methodology that need further consideration, and which could be improved in future work.
Alternating Conditions Design

Alternating conditions designs often begin with a baseline of control conditions (say, No Delay) before other conditions such as (Noncontingent Delay) are introduced in the experimental phase. The purpose of a baseline is to allow time for subjects to adapt to their conditions, and for behaviour levels to stabilise in order to show current levels of performance and suggest future levels. But baseline also entails preexposure to one of the conditions that is involved in the experimental assessment; in the present experiments this was No Delay. Hence, it is unclear whether analyses conducted on data from the experimental phase compare behaviour in Noncontingent Delay (present only during the experimental phase) with (a) No Delay from the experimental phase only, or (b) No Delay from the baseline and experimental phases. Can the experimental conditions be fairly compared when subjects have had greater experience with one of them? Logically, it can be argued that analyses were only based on the behaviour that occurred during the experimental phase, that preexposure or practice would only serve to improve performance in No Delay, and, since Noncontingent Delay performed best anyway, the findings are valid. Perhaps a neater way of overcoming this problem in future investigations would be to omit the baseline and begin with the alternating conditions phase. In this way, potentially confounding factors accompanying preexposure to one condition such as experience, adaptation, practice, and novelty would be distributed equally (theoretically) across conditions.

Reversal

Another difficulty with the present design concerned the reversal to baseline conditions. The return to baseline was introduced to make changes in behaviour levels due to Noncontingent Delay more obvious. It was expected that behaviour levels would be reasonably consistent under No Delay throughout the experiment, but there would be visibly obvious shifts in level, first when Noncontingent Delay was introduced, and later when it was removed. The results were not obvious, however, because behaviour levels under No Delay in the reversal phase were not always similar to No Delay in the previous phases; probably because materials became more difficult as the program progressed. Hence, with hindsight and without a compelling argument to maintain the reversal phase, the sets used in this phase could have been more profitably used for additional comparisons of the experimental conditions.
Internal and External Validity

Learning Phase

For experimental purposes, it would have been advantageous to standardise set length and difficulty, and remove the sequential relationship of sets. This would eliminate any possibility that these factors could covary with the experimental factors, and internal validity would be guaranteed. This was not appropriate in the present thesis, however, because standardisation would not only have taken considerable developmental effort, but would have decreased the external validity of the materials, and an important goal was to work with realistic materials.

Although set size, content, and difficulty did undoubtedly contribute to background "noise", they do not compromise the validity of the findings because they were not systematically associated with the experimental conditions. Confounding was overcome by randomly assigning conditions to sets, and substantially altering this assignment across experiments. The delay intervention was influential enough to be visually and statistically apparent against such background "noise", and regardless of experiment-specific condition-to-set assignment.

Another way of dealing with the potential condition-to-set confound would have been to reverse the order of assignment so that sets (topics) studied under No Delay or Noncontingent Delay in one experiment or subject, were studied under Noncontingent Delay or No Delay, respectively, in another experiment or subject. Alternatively, different condition-to-set assignments could have been used for each of the subjects within one experiment. In either case, experimental control would be demonstrated by obtaining similar delay effects regardless of assignment. In this thesis, identical assignments within each experiment were used so that the results from each subject were directly comparable.

Tests

The test results showed that learning with imposed delays improves retention (Experiments 3 and 4). It is reasonable to assume that because imposed delays promote extra study (Experiment 2) and cause increased accuracy, then this study would also result in greater retention for those materials. But there were two potentially confounding factors that need to be addressed before the findings can be accepted with complete confidence. The first concerns the difficulty of items testing each experimental condition and the second concerns the relative novelty of each condition.
Test item difficulty. To equate test-item difficulty across conditions, the present studies relied upon random assignment of learning sets to conditions, random selection of test items from a large pool, and reasonably large tests. There was, however, no empirical measure of the difficulty of these items, and therefore no empirical measure of the difficulty of the tests ultimately constructed for No Delay and Noncontingent Delay. Without this, there is a possibility that test items assessing Noncontingent Delay were relatively easy, and that this is the reason test scores were best for this condition. A counter argument for this explanation is that Experiments 3 and 4 had quite different test constructions: the test used in Experiment 4 included more items, and, of those items assessing the alternating conditions material, half did not appear in Experiment 3. Nevertheless, the test outcomes were very similar, indicating that it was the experimental variable that was the active component.

Ideally, however, test difficulty should be empirically determined, and one way of accomplishing this is to select test items from the learning sets on the basis of how often students answered them correctly (the method employed by Holland & Porter, 1961). A problem with this method is that it requires many students to have completed the learning sets to provide reliable information about item difficulty. In addition, because this method requires reusing learning frames in a test, problems with practice effects may arise; they may not be systematically related to either experimental condition however. An alternative means of empirically standardising test-item difficulty is to have both naive respondents and those who have worked through the learning sets complete the items. It could then be determined what level of expertise is required to complete each item correctly. Whatever approach is adopted for precise determination of test-item difficulty, an investment of considerable time and effort will be required.

Test size also posed some problems that were related to item difficulty. Because the total test size was 50-60 items, relatively few items assessed each condition (Baseline, No Delay, Noncontingent Delay, Reversal). The problem with this was that small differences in the number of items answered correctly for each condition caused large percentage differences across conditions. The retention advantage of delays was reasonably consistent across subjects and experiments, but it is unclear whether the magnitude of this effect has been reliably established. By equating test item difficulty across conditions, the problem of test size would be largely avoided.
Novelty. A second interpretive difficulty with the present test results concerns the relative novelty of the experimental conditions. Because Noncontingent Delay was introduced after a baseline of No Delay, it was novel. Perhaps novelty increases subjects’ attention, and it was this aspect of the intervention, rather than the programmed delays, that improved retention in Noncontingent Delay. One way of equating the novelty of each condition within an alternating conditions design is to disregard the single-condition baseline and immediately introduce the alternating conditions phase. An alternative is to employ a groups design in which one group of subjects receives delays and the other does not. Because subjects only receive one condition, the relative novelty of each condition is not at issue. It is unfortunate that tests were not employed in Experiment 1 where two equally novel conditions, Noncontingent Delay and Contingent Delay, had different effects on accuracy. If test results corresponded to the accuracy findings (i.e., more delay equalled better accuracy) then the level of delay, and not novelty, would be responsible for this.

Fatigue

Fatigue seemed to be an important factor affecting subjects’ performance. Compared to the error rates expected from programmed materials and the estimated set-completion times provided in the Holland and Skinner text, subjects in Experiments 1 and 2 showed low accuracy and slow times. Possible explanations of this included students being of low ability (or lower than the sample used to establish norms), not receiving response-contingent payment, or being fatigued. In subsequent experiments, similar students and payment schedules were used, but sessions were reduced from three sets to one set. Here, accuracy increased and set times decreased, indicating the importance of session length on performance. In fact, these subjects often had accuracy levels of 80% to 90% even though there was no response-contingent reinforcement. This is a very creditable performance, and indicates that subjects received strong natural reinforcement (e.g., being seen to be intelligent) for good performance. Reducing session length can be procedurally inefficient if it reduces the amount of data that can be obtained in one session, but it appears to be an effective strategy for reducing fatigue and promoting good performance.

Delay Bar

Delay bars are a frequently-employed software feature that inform a user that the computer is busy and indicate how long it will be before interaction can recommence. Perhaps it is this informative dimension of the Delay bar that made subjects in the
present experiments observe the screen during the delay; subjects want to continue as soon as possible, and the Delay bar indicates when this can occur. If delay termination was not signalled by the bar, or if another signal (such as a tone) was used, subjects may not have attended to the screen and benefited from the available learning information. The findings have clearly demonstrated that it is the combination of a brief study period and the materials being present on the screen that improves performance, but the extent to which the Delay bar controlled visual orientation of subjects to the materials is unknown. If it does have this sort of influence, then the Delay bar is analogous to techniques like embedding questions in prose passages, flashing text, animation, or zooming; they are not active ingredients, but they are an important tool for directing observation. If the Delay bar is an orienting tool, and this could be easily determined by comparing delay conditions with and without the bar, or by monitoring eye movements, then its screen position may determine what else is observed during the delay.

Programmed Instruction

As stated at the outset of this thesis, the current experiments were not an assessment of programmed instruction, or other instructional strategies or theoretical positions. However, programmed instruction was the vehicle for assessing the self-pacing tenet, and in light of the interest and controversy it has generated over many years, and suggestions that it is the basis of future success in automated instruction (Vargas & Vargas, 1991), some current observations about programmed instruction are pertinent.

Much of the following information regarding the impressions of subjects in the present experiments was obtained in debriefing sessions when they completed questionnaires (Moore & Smith, 1964; Van Atta, 1961; described in the General Method) and answered follow-up questions of the experimenter.

Easy-Come-Easy-Go Notion

An early criticism of programmed instruction was that because learning did not require a great deal of effort, anything learned would be quickly forgotten; that is, easy come, easy go (see Holland & Porter, 1961; Skinner, 1958). Previous experimental assessments have not supported this criticism (e.g., Holland & Porter, 1961), and neither do the present results. The following information is derived from Experiments 3 to 6 (i.e., those which used tests). Ease of learning the material in the Holland and
Skinner text ("easy come") can be measured in at least two ways: (a) during debriefing subjects rated the difficulty of materials 4 (modal response) on a 7-point scale, indicating that the material was neither too difficult nor too easy (In addition, subjects reported that the step size from frame to frame was appropriate.); and (b) most subjects had high levels of accuracy in learning sets.

One way to determine whether performance declined substantially over time ("easy go") is to assess whether scores in follow-up tests were substantially lower than posttest scores. The single-subject data make this possible. The difference in posttest and follow-up test performance, collapsed across all conditions, was calculated for each subject: Four subjects performed better in the follow-up test than the posttest, six subjects performed better in the posttest than the follow-up test, and for two subjects there was no difference. When there was a difference between test scores, it was typically 5% or less. Hence, there was no reliable decline in performance one-month after the posttest; the materials do not "easily go".

It is notable that test performances were typically high and did not decline over time, because several contingencies were in place during the experiments that were unlikely to promote or sustain optimal levels of performance. First, when subjects sat the posttest they had completed the learning sets and, hence, had earned the majority of their payment (based on attendance, not performance). Completing the two tests did secure their bonus payment (for not missing any sessions), but this was not performance contingent. Second, subjects studied the learning materials without intermittent tests that are used in the standard implementation of the program, and that have been shown to improve retention (Holland & Porter, 1961). Third, the posttest and follow-up test questions were selected from the intermittent tests, so the present subjects completed items out of order, out of context, and up to 1 month after they had studied the relevant material. Finally, in some conditions, subjects also studied without the possible benefits of the review feature. In summary, test performance was generally high and well maintained even though experimental contingencies operated against optimum test performance.

The findings complement those concerning time on task: Procedures that are difficult or that require more time on task do not necessarily improve learning. Properly constructed, programmed materials are one example of an instructional approach that makes learning easy for students, but does not hinder retention.
Affect

The social validity of behavioural interventions is important. Perhaps because of its application to children programmed instruction seems to have fared poorly when its social validity has been assessed. It has been criticised for being boring and dehumanising, but there is ample evidence that this is not necessarily true.

Very positive feedback about instruction from programmed materials was provided by the staff and students participating in the Roanoke experiment (Rushton, 1965). Teachers commented that above average students showed no loss of interest, average students were enthusiastic and worked harder than usual, and below average students were more interested than usual. Student evaluations of their programmed courses showed that most students did consider programs repetitious, but only 24% of students rated them as overly repetitious. Apparently repetition did not necessarily mean unenjoyable. In fact, nearly 50% of respondents considered their programmed course “enjoyable” or “very enjoyable”, about 40% had no strong opinions, and only 12% rated it negatively.

A similar distinction between repetition and dissatisfaction was apparent in the present experiments. When asked to rate the Holland and Skinner program against a hypothetical lecture-based version of the course, subjects rated them equally interesting (mean = mode = 4 on a 7-point scale). Most subjects reported that the programmed nature of the materials was repetitious, but that this was an effective tool for learning. Furthermore, because similar concepts were approached in different ways, this made the materials interesting. Subjects were unanimous in claiming that the program maintained their attention, and most commented that the program’s focus on key issues and reduction of superfluous materials was enjoyable. All subjects recommended the approach for other courses and universities. Students held this positive regard for the program after working with it daily for 1 month. Whether this would change with longer periods of use, or more exposure in other courses is unclear. The subjects did not report that the programmed approach was dehumanising; some would have liked more opportunity to discuss the material with an instructor, but other students reported that they did not miss participating in class discussions. It was amusing that some subjects believed programmed instruction to be a new technique and wondered when it would be more widely used.

In their meta-analysis of studies comparing programmed instruction and traditional instruction, J. A. Kulik, Cohen, and Ebeling (1980) looked for statistically
significant differences in students' ratings of these teaching methods. They found few
studies that reported student evaluations for both methods, but when these data were
reported, there was no difference between the methods in students ratings of course
quality, how much they learned, how much they enjoyed the course, and how hard
they worked.

Whether programmed instruction is boring is unlikely to be an all-or-none issue.
Just as achievement outcomes under programmed instruction probably depend on
various factors (e.g., program quality, entering behaviour of student), so too might
measures of satisfaction be determined by multiple factors such as content matter,
delivery medium, successfulness, and contrast effects with other forms of instruction.

Consistent Feedback

One of the purported advantages to teaching of programmed instruction was an
emphasis on controlling student behaviour by positive, rather than negative,
contingencies of reinforcement. Programmed materials were constructed so that
subjects would often be correct, and would receive consistent and immediate feedback
regarding performance; this was intended to be reinforcing and maintain motivation.
Programmed instruction has been criticised for offering too much feedback: That is, it
is annoying and unnecessary because subjects often are correct and do not need the
continual confirmation.

All of the subjects in the present experiments liked continual feedback because it
provided confirmation when they were correct, and assistance when they were not.
They remarked that it was pleasant (and unusual) to consistently feel that they were on
the right track. Perhaps reports of regular feedback being annoying occurred when it
was used with very easy programs or mechanical teaching machines which were slow
and difficult to manipulate. In contrast, the speed and ease of programming and usage
of computers mean that these potential difficulties should no longer exist.

Future Directions

Electronic and psychological technology will continue to be the two important
components of automated instruction. Electronic media will evolve rapidly and be
technologically exciting, but it is psychological technology that will determine
instructional effectiveness. Some issues regarding the interplay of these two
components are considered now, beginning with the medium and then addressing
learner control.
Automated Instruction

The limited effectiveness of automated instruction probably reflects an immature stage of research into relevant instructional issues (such as learner control). This may be a product of a media assumption that is reluctant to pass away. Because new media are technologically exciting, they have a halo effect, and it is easy to assume that whatever way they deliver instruction, it will be beneficial. But history indicates that this is an erroneous assumption; it is also one that those uninformed of the history of educational media are doomed to repeat (see Pagliaro, 1983).

History documents many episodes in which educational media research has followed a cycle of having a new technology, becoming excited about its teaching potential, making grandiose claims, repeating old research pursuits (e.g., new media vs. traditional media comparison), becoming disappointed by an inability to reliably demonstrate superiority of the new method, and then adopting a subsequent technological advance and repeating the cycle. For example, after Edison invented motion pictures in the early 1900's he predicted that they would revolutionise the educational system and supplant textbooks (from Vargas, 1986). Similar unfulfilled expectations accompanied other media like educational TV, and Pressey's and Skinner's automated devices. Computers also have disappointed as educators. Unlike their predecessors, however, computers maintain a dominant role in schools and society because they are extremely flexible, and can do many noninstructional tasks well (e.g., accounting, desk top publishing). There is not yet, however, an adequate technology of software development, and until there is computers will only have a trivial role in education (De Laurentiis, 1993).

One way in which software can continue to improve is by further incorporating psychological findings. Recently, there has been some signs of a renewed interest in programmed instruction (e.g., Tudor & Bostow, 1991; Vargas & Vargas, 1991). It provides clear foundation principles for designing software (e.g., require plentiful overt responding to enable shaping), is a good base for researching additional reliable principles, and, combined with modern media, could be a very effective tool for the future. It is true that programmed instruction is no longer a fashionable term, but the success of many existing instructional techniques is determined by the extent to which they are programmed (Vargas & Vargas, 1991). These authors also suggest that the technique is still in its infancy and warrants serious reconsideration as an educational
technique for the future. The effectiveness of using programming principles as a basis from which to devise other instructional techniques has been established (e.g., PSI).

**Learner Control**

The more control that learners are given over instructional decisions about pacing, presentation, responding, sequencing, and remediation, the more potential there is for unsophisticated learners to choose poorly and learn less. The relentless push of new technology may exacerbate this problem by providing even more choices. For example, the internet vastly increases the amount of information that is accessible to learners; compact discs and video discs controlled by CAI are becoming common classroom aids that provide random access to video and audio information at extremely fast rates; and multimedia provides many ways of combining and presenting the available information. To date, instructional designers have been largely unsuccessful with prescriptions for learner control. Because emerging technologies increase the number of program dimensions requiring control, the instructional tactic of leaving decisions to the learner may be an easily programmed option, but not one that is instructionally helpful. Control over instructional decisions does not become less important when the number of options increases. There may be more to lose.

Some recent evidence from the newer media (e.g., hypertext) indicates that learners are not benefiting from greater access to information and presentation options. In a recent discussion of self-directed learning approaches (such as hypertext), Reader (1993) maintained that studies of these media show that students are failing to access information in a goal-directed manner and be active in learning the material. This appears to be evidence of technology displacing psychology, because the necessity to include such study behaviours is long established. Clear objectives, structured materials, and active participation are trademarks of behaviourally-based instructional approaches.

In addition, a meta-analysis of learning outcomes from studying with interactive video found that (a) it led to smaller effect sizes than traditional instruction, and (b) student control over instructional path or remediation decreased outcomes compared to guided (program) control (McNeil & Nelson, 1991). The researchers concluded that instruction via interactive video is probably best accomplished when control is not left entirely with the learner. They also reiterated a point made earlier in this thesis, that (interactive video) research may benefit by a shift from comparative studies to a focus on features that increase learning.
Individualised instruction is gaining increasing emphasis and is appearing in different guises (e.g., "distance" education) as technology provides greater access to information. Access does not equate with learning however, and further consideration of individualised instructional approaches is warranted. The burden of responsibility regarding fundamental control issues (e.g., self-pacing tenet) lies not with the learner, but with those who design their learning environment; instructional designers, psychologists, and behaviour analysts.
APPENDICES

Appendix 2-1:
Table of contents from *The Analysis of Behavior* (Holland & Skinner, 1961)

<table>
<thead>
<tr>
<th>Part</th>
<th>REFLEX BEHAVIOR</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 1</td>
<td>Simple reflexes</td>
<td>54</td>
<td>1</td>
</tr>
<tr>
<td>Set 2</td>
<td>Conditioned reflexes</td>
<td>30</td>
<td>9</td>
</tr>
<tr>
<td>Set 3</td>
<td>Conditioned reflexes (continued)</td>
<td>29</td>
<td>14</td>
</tr>
<tr>
<td>Set 4</td>
<td>Pavlov's experiments</td>
<td>29</td>
<td>20</td>
</tr>
<tr>
<td>Set 5</td>
<td>Conditioned reflexes (continued)</td>
<td>59</td>
<td>26</td>
</tr>
<tr>
<td>Set 6</td>
<td>Response mechanisms</td>
<td>29</td>
<td>35</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part II</th>
<th>OPERANT CONDITIONING: ELEMENTARY CONCEPTS</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 7</td>
<td>Introduction to operant conditioning</td>
<td>30</td>
<td>41</td>
</tr>
<tr>
<td>Set 8</td>
<td>The standard experimental situation</td>
<td>31</td>
<td>46</td>
</tr>
<tr>
<td>Set 9</td>
<td>Positive and negative reinforcement</td>
<td>29</td>
<td>52</td>
</tr>
<tr>
<td>Set 10</td>
<td>Basic concepts applied</td>
<td>32</td>
<td>57</td>
</tr>
<tr>
<td>Set 11</td>
<td>Conditioned reinforcers</td>
<td>60</td>
<td>62</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part III</th>
<th>OPERANT CONDITIONING: PRECISE CONTINGENCIES</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 12</td>
<td>The cumulative recorder</td>
<td>30</td>
<td>72</td>
</tr>
<tr>
<td>Set 13</td>
<td>Factors affecting speed of conditioning</td>
<td>61</td>
<td>78</td>
</tr>
<tr>
<td>Set 14</td>
<td>Accidental contingencies and superstitious behavior</td>
<td>51</td>
<td>88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Part IV</th>
<th>SHAPING</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Set 15</td>
<td>Principles of shaping new behavior</td>
<td>48</td>
<td>97</td>
</tr>
<tr>
<td>Set 16</td>
<td>Applications of principles of shaping</td>
<td>29</td>
<td>106</td>
</tr>
</tbody>
</table>

| Review  | Test covering parts I-IV                             | 39     | 111  |
### Part V: INTERMITTENT REINFORCEMENT

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>Schedules defined: Fixed-interval schedules</td>
<td>40</td>
<td>117</td>
</tr>
<tr>
<td>19</td>
<td>Variable-interval, fixed-ratio, and variable-ratio schedules</td>
<td>47</td>
<td>124</td>
</tr>
<tr>
<td>20</td>
<td>Schedules of reinforcement: Summary and review</td>
<td>33</td>
<td>132</td>
</tr>
</tbody>
</table>

### Part VI: STIMULUS CONTROL

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>Stimulus discrimination</td>
<td>75</td>
<td>137</td>
</tr>
<tr>
<td>22</td>
<td>Stimulus generalization</td>
<td>71</td>
<td>149</td>
</tr>
<tr>
<td>23</td>
<td>Chaining</td>
<td>37</td>
<td>161</td>
</tr>
<tr>
<td>24</td>
<td>Shaping continuous repertoires</td>
<td>42</td>
<td>167</td>
</tr>
<tr>
<td>25</td>
<td>Discrete and continuous repertoires</td>
<td>37</td>
<td>175</td>
</tr>
</tbody>
</table>

### Part VII: DEPRIVATION

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>26</td>
<td>Basic concepts</td>
<td>68</td>
<td>181</td>
</tr>
<tr>
<td>27</td>
<td>Generalized reinforcers</td>
<td>33</td>
<td>192</td>
</tr>
<tr>
<td>28</td>
<td>Feeding cycles</td>
<td>31</td>
<td>197</td>
</tr>
</tbody>
</table>

### Review

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>Test covering Parts V-VII</td>
<td>28</td>
<td>203</td>
</tr>
</tbody>
</table>

### Part VIII: EMOTION I

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Activation syndrome</td>
<td>32</td>
<td>208</td>
</tr>
<tr>
<td>31</td>
<td>Predispositions in emotion</td>
<td>32</td>
<td>213</td>
</tr>
</tbody>
</table>

### Part IX: AVOIDANCE AND ESCAPE BEHAVIOR

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>32</td>
<td>Basic concepts</td>
<td>39</td>
<td>218</td>
</tr>
<tr>
<td>33</td>
<td>Analysis of examples of avoidance and escape</td>
<td>30</td>
<td>224</td>
</tr>
<tr>
<td>34</td>
<td>Avoidance experiments</td>
<td>31</td>
<td>229</td>
</tr>
</tbody>
</table>

### Part X: EMOTION

<table>
<thead>
<tr>
<th>Set</th>
<th>Title</th>
<th>Frames</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>Experiments on anxiety</td>
<td>30</td>
<td>235</td>
</tr>
<tr>
<td>36</td>
<td>Emotions as reinforcing and aversive conditions</td>
<td>25</td>
<td>241</td>
</tr>
</tbody>
</table>
Part XI  PUNISHMENT
Set 37  Basic concepts  32  245
Set 38  Effects of punishment during extinction of reinforced behavior  30  251
Set 39  Additional effects of punishment, functions of the aversive stimulus  31  257
Set 40  Effects of continuous punishment  40  262

Review
Set 41  Test covering parts VIII-XI  23  272

Part XII  SCIENTIFIC ANALYSIS AND THE INTERPRETATION OF COMPLEX CASES
Set 42  Goals and techniques of science  33  276
Set 43  Multiple effects  29  281
Set 44  Multiple causes and conflicting responses  33  287
Set 45  A problem in behavioral engineering  32  292

Part XIII  SELF-CONTROL
Set 46  Analysis of voluntary and involuntary behavior  31  298
Set 47  Techniques of self-control  33  303

Part XIV  INTERPRETATION OF PERSONALITY
Set 48  Inadequate self-knowledge  33  308
Set 49  Rationalization  30  313
Set 50  Drug addiction  27  318
Set 51  Aggressiveness, withdrawal, and reaction formation  31  323
Set 52  Psychotherapy  28  328

Review
Set 53  Test covering parts XII-XIV  27  333
Appendix 2-2: Subject details

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Subject</th>
<th>Gender</th>
<th>Age</th>
<th>Year of study</th>
<th>School (major)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H21</td>
<td>F</td>
<td>30</td>
<td>3</td>
<td>B.A. (Humanities)</td>
</tr>
<tr>
<td></td>
<td>H22</td>
<td>F</td>
<td>23</td>
<td>3</td>
<td>B.A. (Psychology)</td>
</tr>
<tr>
<td></td>
<td>H23</td>
<td>F</td>
<td>23</td>
<td>3</td>
<td>B.A. (Social Sciences)</td>
</tr>
<tr>
<td></td>
<td>H24</td>
<td>F</td>
<td>20</td>
<td>3</td>
<td>B.A. (Humanities)</td>
</tr>
<tr>
<td>2</td>
<td>H25</td>
<td>F</td>
<td>18</td>
<td>1</td>
<td>B.N. (Nursing)</td>
</tr>
<tr>
<td></td>
<td>H26</td>
<td>F</td>
<td>19</td>
<td>1</td>
<td>B.N. (Nursing)</td>
</tr>
<tr>
<td></td>
<td>H27</td>
<td>F</td>
<td>19</td>
<td>2</td>
<td>B.A. (Psychology)</td>
</tr>
<tr>
<td></td>
<td>H28</td>
<td>M</td>
<td>18</td>
<td>1</td>
<td>Dip. (Tertiary Studies)</td>
</tr>
<tr>
<td>3</td>
<td>H35</td>
<td>F</td>
<td>22</td>
<td>5</td>
<td>B.Sc. (Chemistry)</td>
</tr>
<tr>
<td></td>
<td>H36</td>
<td>F</td>
<td>21</td>
<td>3</td>
<td>B.Sc. (Biology)</td>
</tr>
<tr>
<td></td>
<td>H37</td>
<td>M</td>
<td>19</td>
<td>2</td>
<td>B.Sc. (Mathematics)</td>
</tr>
<tr>
<td>4</td>
<td>H44</td>
<td>M</td>
<td>21</td>
<td>4</td>
<td>B.Sc./B.Comm.</td>
</tr>
<tr>
<td></td>
<td>H45</td>
<td>F</td>
<td>19</td>
<td>2</td>
<td>B.Sc. (Comp./Maths)</td>
</tr>
<tr>
<td></td>
<td>H46</td>
<td>F</td>
<td>21</td>
<td>3</td>
<td>B.Sc. (Biology)</td>
</tr>
<tr>
<td>5</td>
<td>H38</td>
<td>F</td>
<td>19</td>
<td>2</td>
<td>B.Sc. (Biology)</td>
</tr>
<tr>
<td></td>
<td>H39</td>
<td>M</td>
<td>29</td>
<td>2</td>
<td>B.Sc. (Biology)</td>
</tr>
<tr>
<td></td>
<td>H40</td>
<td>F</td>
<td>20</td>
<td>2</td>
<td>B.Sc. (Biochemistry)</td>
</tr>
<tr>
<td>6</td>
<td>H41</td>
<td>F</td>
<td>21</td>
<td>3</td>
<td>B.A. (Literature)</td>
</tr>
<tr>
<td></td>
<td>H42</td>
<td>M</td>
<td>25</td>
<td>3</td>
<td>B.A. (Architecture)</td>
</tr>
<tr>
<td></td>
<td>H43</td>
<td>F</td>
<td>20</td>
<td>3</td>
<td>B.Sc. (Biology)</td>
</tr>
</tbody>
</table>
Appendix 2-3: Experimental instructions for Experiment 2

In this experiment you will use computer-assisted instruction to learn about an important area of Psychology: the experimental analysis of behaviour and its application. You will cover three topics per session, and there will be one session every day (Monday to Friday) for 15 sessions; each session will last between 1 and $1\frac{1}{2}$ hours.

Each topic has between 25 and 75 questions and for each you will supply one or two missing words. This study has three experimental conditions: (1) no delay between questions, (2) a 10-second delay after each question, and (3) a 10-second delay plus blank screen between all questions. During a lesson only one of these conditions will apply (shown in yellow at the top of the screen). For the first five sessions you will receive only Condition 1 but for all other sessions you will receive all three conditions. You must answer all questions correctly before you finish a topic, so when you have answered the final question, the program will take you back to the questions you answered incorrectly, and will continue this process until all questions have been answered correctly. At the end of each lesson, you will be asked a question concerning the lesson, and for this you will type a number between 1 and 9 to show your answer.

Between sessions do not study the material or talk to other people about the topics; this might ruin the experiment, and make the results uninterpretable.
Appendix 2-4: Experimental instructions for Experiments 3 and 4

In this experiment you will use computer-assisted instruction to learn about an important area of Psychology: the experimental analysis of behaviour and its application.

Initially, you will complete a brief test to determine your familiarity with this topic. You are likely to find that answering these questions correctly is difficult (or impossible!). Do not despair, no prior knowledge is expected. You will then cover one topic per session, and there will be two sessions (one morning, and one afternoon) every day (Monday to Friday) for 40 sessions; each of these will last between 30 and 40 minutes.

Each topic has between 25 and 75 questions and for each you will supply one or two missing words. This study has two experimental conditions, they differ only in the amount of delay between questions: (1) no delay between questions, and (2) a 10-second delay between all questions. During a lesson only one of these conditions will apply (shown in yellow at the top of the screen) but throughout the experiment you will receive both conditions.

You must answer all questions correctly before you finish a topic, so when you have answered the final question, the program will take you back to the questions you answered incorrectly, and will continue this process until all questions have been answered correctly. At the end of each lesson, you will be asked a question concerning the lesson, and for this you will type a number between 1 and 9 to show your answer.

After finishing all the topics you will complete two review tests, one immediately, and the second about one month later. Until the experiment is complete do not study the material or talk to other people about the topics; this might ruin the experiment, and make the results uninterpretable.

The instructions for Experiments 3 and 4 were identical except that the first sentence of the fourth paragraph (referring to the Review) was removed for Experiment 4.
Appendix 2-5: Experimental instructions for Experiments 5 and 6

In this experiment you will use computer-assisted instruction to learn about an important area of Psychology: the experimental analysis of behaviour and its application.

Initially, you will complete a brief test to determine your familiarity with this topic. You are likely to find that answering these questions correctly is difficult (or impossible!). Do not despair, no prior knowledge is expected. You will then cover 1 topic per session, and there will be two sessions (one morning, and one afternoon) every day (Monday to Friday) for 40 sessions; each of these will last between 30 and 40 minutes.

Each topic has between 25 and 75 questions and for each you will supply one or two missing words. This study has two experimental conditions which differ only in the presence of a “Review”: (1) when a Review exists you must answer all questions correctly before you finish a topic; so when you have answered the final question, the program will take you back to the questions you answered incorrectly, and will continue this process until all questions have been answered correctly; and (2) no Review. During a lesson only one of these conditions will apply (shown in yellow at the top of the screen) but throughout the experiment you will receive both conditions.

At the end of each lesson, you will be asked a question concerning the lesson, and for this you will type a number between 1 and 9 to show your answer. After finishing all the topics you will complete two review tests, one immediately, and the second about one month later.

Until the experiment is complete do not study the material or talk to other people about the topics; this might ruin the experiment, and make the results uninterpretable.
Appendix 2-6: Consent form for Experiments 1 and 2

The experimental procedures of Glenn Kelly's study have been explained to me, and I hereby agree to participate in this study on the condition that I may cease my involvement at any time. I understand that I will receive $10 for every session I complete, and that if I participate in all scheduled sessions (maximum of 15) I will receive a bonus of $25. I also agree that I shall lose $10 for every scheduled session that I miss.

Signed ........................................................................

Date ..................................................

I also agree to the above-mentioned conditions.

Signed ........................................................................

Date ..................................................

Demographic details

Name ........................................................................

Address ........................................................................

Phone ..................................................

Age (in years) ..................................................

Gender (M or F) ..................................................

Year of study (e.g., 1, 2, 3, 4) ...............

Major programme ............................................................

aThe consent forms for Experiments 1 and 2 were identical except that information regarding payment was absent for Experiment 1, because subjects participated as part of the course requirements of a third-year research project.
Appendix 2-7: Consent form for Experiments 3 to 6

The experimental procedures of Glenn Kelly’s study have been explained to me, and I hereby agree to participate in this study on the condition that I may cease my involvement at any time. I understand that I will receive $5 for every session I complete, and that if I participate in all scheduled sessions (Maximum of 43) I will receive a bonus of $50. I also agree that I shall lose $15 for every scheduled session that I miss. I understand that upon completion of the post-test I will be paid half of the money owed to me, and that the remaining money will be paid upon completion of the follow-up test (about 1 month later). Furthermore, I permit the use of my academic record for analytical purposes provided that this information will remain totally anonymous.

Signed .........................................................................................

Date ..................................................

I also agree to the above-mentioned conditions.

Signed .........................................................................................

Date ..................................................

Demographic details

Name ...........................................................................................

Address ..........................................................................................

Phone .................................................................

Age (in years) .........................

Gender (M or F) .........................

Year of study (e.g., 1, 2, 3, 4) ...........

Major programme ..........................
Appendix 2-8: Example screen display

Set 1: Simple Reflexes

Question 1 of 54

No delay between questions

A doctor taps your knee (patellar tendon) with a rubber hammer to test your ________

Your Answer:

Please type your answer
Set 1: Simple Reflexes

Question 1 of 54

No delay between questions

A doctor taps your knee (patellar tendon) with a rubber hammer to test your

Your Answer: reflexes

Please type your answer
A doctor taps your knee (patellar tendon) with a rubber hammer to test your reflexes.

No delay between questions

Your Answer: reflexes
Right Answer: reflexes (reflex)

Press C if your answer was correct or 1 if it was incorrect.
Set 1: Simple Reflexes

Question 1 of 54

No delay between questions

A doctor taps your knee (patellar tendon) with a rubber hammer to test your ______.

Your Answer: reflexes
Right Answer: reflexes (reflex)

Press Esc to stop, or any other key to continue
Appendix 2-12:
Screen display in contingent-delay and noncontingent-delay conditions

Incorrect 0
Correct 1

A doctor taps your knee (patellar tendon) with a rubber hammer to test your _____

Set 1: Simple Reflexes
Question 1 of 54

A delay before each question

Your Answer: reflexes
Right Answer: reflexes (reflex)

Please wait until the bar is gone
Appendices

Appendix 2-13: Percentage of incorrectly scored responses

<table>
<thead>
<tr>
<th>Experiment</th>
<th>Subject</th>
<th>Type A Error</th>
<th>Type B Error</th>
<th>Total Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>H21</td>
<td>1.94</td>
<td>0.00</td>
<td>1.94</td>
</tr>
<tr>
<td></td>
<td>H22</td>
<td>0.37</td>
<td>0.05</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>H23</td>
<td>1.51</td>
<td>0.00</td>
<td>1.51</td>
</tr>
<tr>
<td></td>
<td>H24</td>
<td>1.50</td>
<td>0.00</td>
<td>1.50</td>
</tr>
<tr>
<td></td>
<td>H25</td>
<td>1.85</td>
<td>0.10</td>
<td>1.95</td>
</tr>
<tr>
<td>2</td>
<td>H26</td>
<td>2.06</td>
<td>0.04</td>
<td>2.10</td>
</tr>
<tr>
<td></td>
<td>H27</td>
<td>1.34</td>
<td>0.27</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>H28</td>
<td>4.84</td>
<td>0.00</td>
<td>4.84</td>
</tr>
<tr>
<td>3</td>
<td>H35</td>
<td>0.47</td>
<td>0.11</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>H36</td>
<td>0.95</td>
<td>0.16</td>
<td>1.11</td>
</tr>
<tr>
<td></td>
<td>H37</td>
<td>1.06</td>
<td>0.17</td>
<td>1.22</td>
</tr>
<tr>
<td>4</td>
<td>H44</td>
<td>0.71</td>
<td>0.26</td>
<td>0.96*</td>
</tr>
<tr>
<td></td>
<td>H45</td>
<td>1.28</td>
<td>0.26</td>
<td>1.54</td>
</tr>
<tr>
<td></td>
<td>H46</td>
<td>2.05</td>
<td>0.64</td>
<td>2.69</td>
</tr>
<tr>
<td>5</td>
<td>H38</td>
<td>0.35</td>
<td>0.35</td>
<td>0.69*</td>
</tr>
<tr>
<td></td>
<td>H39</td>
<td>1.86</td>
<td>0.06</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>H40</td>
<td>0.23</td>
<td>0.46</td>
<td>0.70*</td>
</tr>
<tr>
<td>6</td>
<td>H41</td>
<td>0.11</td>
<td>1.50</td>
<td>1.61</td>
</tr>
<tr>
<td></td>
<td>H42</td>
<td>4.02</td>
<td>0.12</td>
<td>4.14</td>
</tr>
<tr>
<td></td>
<td>H43</td>
<td>0.06</td>
<td>0.61</td>
<td>0.67</td>
</tr>
</tbody>
</table>

**Note.** The guideline used for assessing correctness of responses was adopted from Holland and Skinner (1961), and is presented in the General Method. Type A errors occur when an incorrect response is scored as correct, and Type B errors occur when correct responses are scored as incorrect. * Discrepancies between part scores and their total is due to rounding.
Appendix 3-1: Sequence of conditions for the subjects in Experiment 1

<table>
<thead>
<tr>
<th>Phase</th>
<th>Session</th>
<th>Ordinal Position Within Session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>2</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Baseline</td>
<td>3</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>CD</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>CD</td>
</tr>
<tr>
<td>Alternating Conditions</td>
<td>10</td>
<td>ND</td>
</tr>
<tr>
<td>11</td>
<td>NCD</td>
<td>ND</td>
</tr>
<tr>
<td>12</td>
<td>CD</td>
<td>ND</td>
</tr>
<tr>
<td>13</td>
<td>ND</td>
<td>NCD</td>
</tr>
<tr>
<td>14</td>
<td>NCD</td>
<td>CD</td>
</tr>
<tr>
<td>15</td>
<td>CD</td>
<td>NCD</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay, and CD = Contingent Delay.

aFor Experiment 2, the assignment of ND and NCD remains unchanged but CD is replaced by NCB.
### Appendix 4-1: Sequence of conditions presented to subjects in Experiment 3

<table>
<thead>
<tr>
<th>Phase</th>
<th>Sessions</th>
<th>Condition During Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1-10</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>11-12</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>13-14</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>15-16</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>17-18</td>
<td>NCD</td>
</tr>
<tr>
<td>Alternating Conditions</td>
<td>19-20</td>
<td>ND</td>
</tr>
<tr>
<td>Conditions</td>
<td>21-22</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>23-24</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>25-26</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>27-28</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>29-30</td>
<td>ND</td>
</tr>
<tr>
<td>Reversal</td>
<td>31-40</td>
<td>ND</td>
</tr>
</tbody>
</table>

**Note.** ND = No Delay, NCD = Noncontingent Delay.
Appendix 4-2: Sequence of conditions presented to subjects in Experiment 4

<table>
<thead>
<tr>
<th>Phase</th>
<th>Sessions</th>
<th>Condition During Session</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1-10</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>11-12</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>13-14</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>15-16</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>17-18</td>
<td>NCD</td>
</tr>
<tr>
<td>Alternating</td>
<td>19-20</td>
<td>ND</td>
</tr>
<tr>
<td>Conditions</td>
<td>21-22</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>23-24</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>25-26</td>
<td>NCD</td>
</tr>
<tr>
<td></td>
<td>27-28</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>29-30</td>
<td>ND</td>
</tr>
<tr>
<td>Reversal</td>
<td>31-40</td>
<td>ND</td>
</tr>
</tbody>
</table>

Note. ND = No Delay, NCD = Noncontingent Delay.
Appendix 5-1:
Sequence of conditions presented to subjects in Experiments 5 and 6

<table>
<thead>
<tr>
<th>Phase</th>
<th>Sessions</th>
<th>Condition During Session</th>
<th>Experiment 5</th>
<th>Experiment 6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>1-10</td>
<td>REV</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>11-12</td>
<td>REV</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>13-14</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td></td>
<td>15-16</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td></td>
<td>17-18</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td>Alternating</td>
<td>19-20</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td>Conditions</td>
<td>21-22</td>
<td>NR</td>
<td>REV</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>23-24</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td></td>
<td>25-26</td>
<td>NR</td>
<td>REV</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>27-28</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td></td>
<td>29-30</td>
<td>REV</td>
<td>NR</td>
<td>REV</td>
</tr>
<tr>
<td>Reversal</td>
<td>31-40</td>
<td>REV</td>
<td>NR</td>
<td></td>
</tr>
</tbody>
</table>

Note. REV = Review, NR = No Review.
REFERENCES


References


Ice cream for the right answers. (1968, August 1). *Forbes*, p. 46.


References


Merrill, M. D. (1988). Don't bother me with instructional design - I'm busy programming!: Suggestions for more effective educational software. Computers in Human Behavior, 4, 37-52.


References


References


The teaching machines (1960). Time, 76, 91-92, 94.
