Spacing Effect in Memory: Data, Model and Parameters

Philip Pavlik & John Anderson
Carnegie Mellon University
ACT-R Workshop 2003
Spacing Effect

- Advantage of wide spacing of practice vs. narrow spacing of practice
  - Wide spacing of practice generally results in more **durable** learning
  - Ubiquitous result
  - Generally **all memory paradigms** can be shown to demonstrate a spacing effect
  - Can be shown across **various time spans**
  - Occurs in lower **animals and single neurons**
  - AKA **distributed vs. massed practice**
Experiment

Continuous paired-associates

- Initial study opportunity for an English-Japanese pair followed by a sequence of test presentations at different spacings.
- Levels of spacing and practice were manipulated within-subjects and distributed over a first session.
- Performance was measured on a second session.
- “Correct Test” or “Failure and Study” for each presentation
  - Allows modeling each condition as a constant number of presentations
- 2 session experiment with 1 or 7 days between sessions
  - 3 levels of spacing – 2, 14 or 98 intervening trials spacing on session 1
  - 4 levels of practice – 1, 2, 4, or 8 tests on session 1
  - 3 x 4 design within subjects
    - All trials mixed so that intervening trials were either filler items or items from other conditions
    - Session 2 assessed effects of session 1 conditions
      - 4 more practices at 98 spacing for each condition
Quantitative Results: Session 1

- Includes 1, 2, 4 and 8 test conditions
- Includes 1, 2, 4 and 8 test conditions
- Includes 1, 2, and 4 test conditions
- 2 SE bars
Quantitative Results: Session 2

2 presentations (1 + 1) on session 1

3 presentations (1 + 2) on session 1

Aggregated for both retention conditions (1 and 7 days)
Quantitative Results: Session 2

5 presentations (1 + 4) on session 1

9 presentations (1 + 8) on session 1

Aggregated for both retention conditions
Graph of Crossover Interaction

• Strong indication that forgetting is less after wider spacing
  – Average of session 1 final tests for 1, 2 and 4 test conditions
  – Average of session 2 initial tests for 1, 2 and 4 test conditions

Not a different fit… Just a different graph
The Model Needed 2 New Mechanisms

• For Spacing
  – Similar mechanism to Anderson and Schooler (1991)
    • $\text{decay} = f(\text{time from previous presentation})$
    • Wider spacing $\to$ less decay
  – New mechanism proposes that higher activation results in greater decay
    • $\text{decay} = f(\text{activation})$
    • Wider spacing $\to$ less activation $\to$ less decay

• For Long-term forgetting
  – We used (Anderson, Fincham and Douglas, 1999) solution but will not detail this today. But it worked pretty well.
For the Curious: Fit of Current Base Level Equation to the crossover interaction

- Basically the fit degenerates

Includes slowed clock assumption
Decay Equation

\[ d_j = ce^{m_j} + a \quad \Rightarrow \quad m_j = \ln \left( \sum_{j=1}^{n} t_j^{-d_j} \right) \]

- Captures the intuition that there should be
  - diminishing marginal returns for the long-term effect of practice as practice accumulates (the 20th practice on the same session should add less to long-term memory than the 1st)
- Results in the spacing effect
  - more spacing \(\rightarrow\) more forgetting \(\rightarrow\) less activation \(\rightarrow\) less decay \(\rightarrow\) better retention
- Better fits than Anderson and Schooler (1991)
  - Also Anderson and Schooler (1991) tends to over-predict performance with large amounts of practice
Now

\( d_1 = a \)
\( m = \ln(9^{\cdot d_1}) \)
\( d_2 = a + c e^m \)

Now

\( d_3 = a + c e^m \)

\( m = \ln(24^{-d_1} + 15^{-d_2}) \)

- The decay for each presentation is set at the time of the presentation.
- That decay is used in subsequent calculations.

DEMO
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>decay intercept $a$</td>
<td>.170</td>
<td>.208</td>
<td>.120</td>
<td>.143</td>
<td>.294</td>
<td>.058</td>
</tr>
<tr>
<td>decay scale $c$</td>
<td>.219</td>
<td>.138</td>
<td>.366</td>
<td>.488</td>
<td>.415</td>
<td>.276</td>
</tr>
<tr>
<td>threshold $(\tau)$</td>
<td>-.679</td>
<td>-.679*</td>
<td>-.679*</td>
<td>-.679*</td>
<td>-.679*</td>
<td>-.679*</td>
</tr>
<tr>
<td>noise $s$</td>
<td>.255</td>
<td>.255*</td>
<td>.255*</td>
<td>.255*</td>
<td>.255*</td>
<td>.255*</td>
</tr>
<tr>
<td>h scalar $h$</td>
<td>.031</td>
<td>.031*</td>
<td>.031*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>encoding $(b)$</td>
<td>3.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>encoding $(b_r)$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fit Statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$r^2$</td>
<td>.944</td>
<td>.929</td>
<td>.940</td>
<td>.925</td>
<td>.461</td>
<td>.944</td>
</tr>
<tr>
<td>RMSD adjusted</td>
<td>.045</td>
<td>.058</td>
<td>.052</td>
<td>.021</td>
<td>.025</td>
<td>.026</td>
</tr>
<tr>
<td>$\chi^2$</td>
<td>322</td>
<td>246</td>
<td>90.7</td>
<td>42.1</td>
<td>8.77</td>
<td>31.8</td>
</tr>
<tr>
<td>$\chi^2 df$</td>
<td>157</td>
<td>27</td>
<td>40</td>
<td>38</td>
<td>16</td>
<td>20</td>
</tr>
</tbody>
</table>
Do I really need this?

- But isn’t .5 good enough for many applications?
  - Often it is completely adequate....
  - But I’d be curious to hear your results if you decide to let decay vary and fix tau and s at the defaults I found.

- Under what conditions can I ignore this variable decay mechanism safely?
  - If retrievals don’t occur after longer delays
    - Since effects of different decays are greater with more time
    - How long is long depends on the stimuli as we will see
  - If there isn’t too much variability in the amount of practice or spacing of practice for memory items in the model
    - Regular equations are somewhat flexible to different amounts of practice
    - Regular equations are least flexible to comparisons involving different levels of spacing
What does fixing tau and s imply in this model?

- It indicates that the level of performance and variability of performance can often be captured with only the decay parameters.
- However, we do not claim that threshold is necessarily fixed.
However, s may be fixed

- When s is fixed two things happen:
  - Decay parameters are estimated that model the variability of activation during learning
    - Higher or lower decay results in more or less fluctuation of absolute activation across a series of practices
  - Threshold parameter finds a value to capture the absolute average performance
- So, when using this model, it should be unnecessary to estimate s, leave it fixed at .25
- Therefore this mechanism could be accused of adding at most 1 free parameter to most models
  - Since -.7 appears to work for tau we could suggest the new mechanism adds 0 free parameters to most models
  - However, using spreading activation or partial matching would likely result in the need to estimate tau.
    - However, tau would likely be stable across similar models
  - Different memory paradigms (recognition vs recall) may show different taus.
What does a decay function look like?

- Graph below depicts decay for a presentation as a function of activation at the time of that presentation.
- Appears that decay parameters reflect differences in stimuli.
- Error bars reflect .5 SD since activation is a noisy value.

Meaningful Stimuli (words) = low decay
Meaningless Stimuli (CVCs, CCCs and numbers) = high decay
More specifically the slope of the lines (defined by c parameter) may represent resistance to additional practice caused by stimuli or task associations.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>decay intercept a</td>
<td>.170</td>
<td>.208</td>
<td>.058</td>
<td>.143</td>
<td>.294</td>
<td>.120</td>
</tr>
<tr>
<td>decay scale c</td>
<td>.219</td>
<td>.138</td>
<td>.276</td>
<td>.488</td>
<td>.415</td>
<td>.366</td>
</tr>
</tbody>
</table>

**Forced Grouping (2 levels of c and best fitting a)**

- **More varied associations**
  - These experiments had meaningful stimuli and intentional learning
  - These experiments had meaningless stimuli or incidental learning

**Fit Statistics**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$\chi^2$ old</td>
<td>322</td>
<td>246</td>
<td>31.8</td>
<td>42.1</td>
<td>8.77</td>
<td>90.7</td>
</tr>
<tr>
<td>$\chi^2$ new</td>
<td>391</td>
<td>328</td>
<td>32.6</td>
<td>48.2</td>
<td>8.80</td>
<td>104</td>
</tr>
<tr>
<td>$\chi^2$ df</td>
<td>157</td>
<td>27</td>
<td>20</td>
<td>38</td>
<td>16</td>
<td>40</td>
</tr>
</tbody>
</table>
Contributions of the Model

- Additional practical power of ACT-R
  - Now handles spacing well
  - Backwards compatible
    - Can be turned off by setting c=0 so using this mechanism would not be a big commitment if it was available in ACT-R
    - Implies .5 is not a bad estimate of decay
  - Since other parameters can be fixed
    - Implies only 0 or 1 new free parameters
    - Instead of finding best tau and s, find best a and c
      - Or perhaps tau, a, and c.

- Additional theoretical power for ACT-R
  - Wider range of phenomena can be modeled
  - Relatively parsimonious with converging evidence
  - Appears to result in meaningful parameters
Where do we fit in with other theories of Spacing Effect

- **Variable Encoding** (Martin, 1968)
  - More varied contextual elements encoded for spaced trials

- **Voluntary Encoding/Rehearsal** (Atkinson and Shiffrin, 1968; Rundus, 1971)
  - More rehearsals/time in short-term memory for spaced trials

- **Consolidation** (Peterson, 1966; Landauer, 1969)
  - More time for the last trial to be consolidated for spaced trials

- **Effort/Difficulty** (Bjork & Whitten, 1977; Cuddy & Jacoby, 1982)
  - Greater difficulty of spaced trials leads to fuller processing and better memory

- **Habituation** (Hintzman, 1974)
  - Priming of the item results in decreased processing of massed trials

- **Forgetting** (Wickelgren, 1973; Anderson & Schooler; 1991)
  - Wider spacing leads to less forgetting
## Where does the effect occur?

<table>
<thead>
<tr>
<th>Locus of Effect</th>
<th>Voluntary</th>
<th>Involuntary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between P1-P2</td>
<td>Rundus (1971) <em>Rehearsal</em></td>
<td>Landauer (1969), Atkinson and Shiffrin (1968) <em>Consolidation or STS transfer</em></td>
</tr>
</tbody>
</table>
Retention Intervals

- Model 2
- Model 8
- Model 32
- Model 64
- Humans 2
- Humans 8
- Humans 32
- Humans 64

Spacing interval (trials)

Probability Correct
Spacing interval (trials) vs. Probability Correct.

- Dotted line with diamonds represents the Model.
- Solid line with squares represents Humans.

The graph shows the probability of correct responses for both the model and humans across different spacing intervals.