Spacing Effect in Memory: Data, Model and Parameters

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

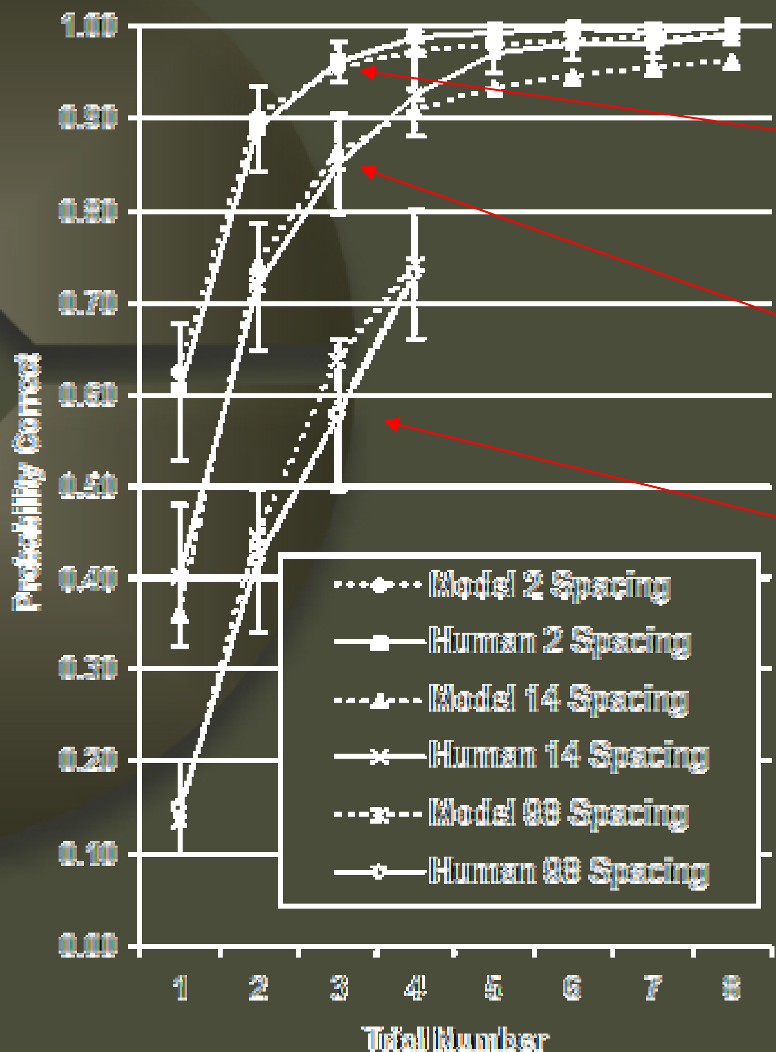
Spacing Effect

- 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

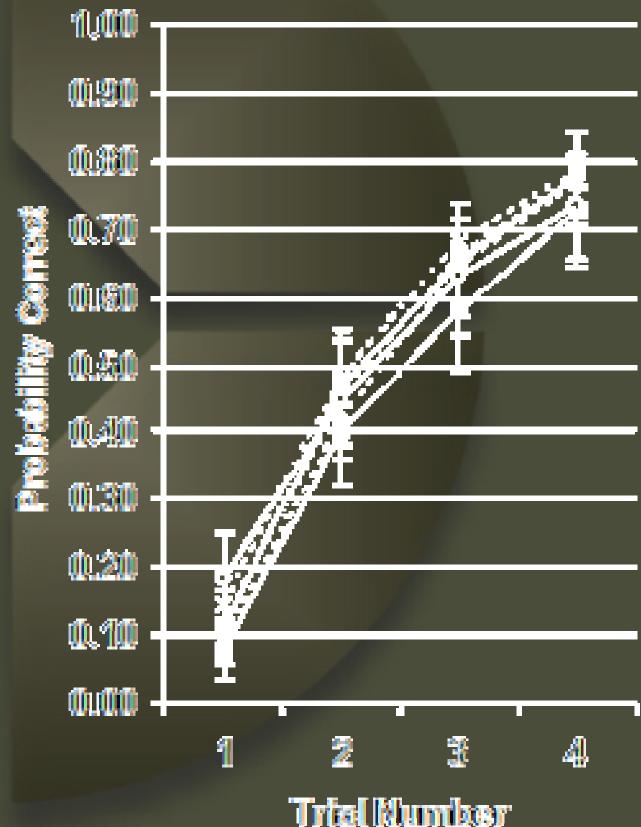
Includes 1,2,4 and 8 test conditions

Includes 1,2, and 4 test conditions

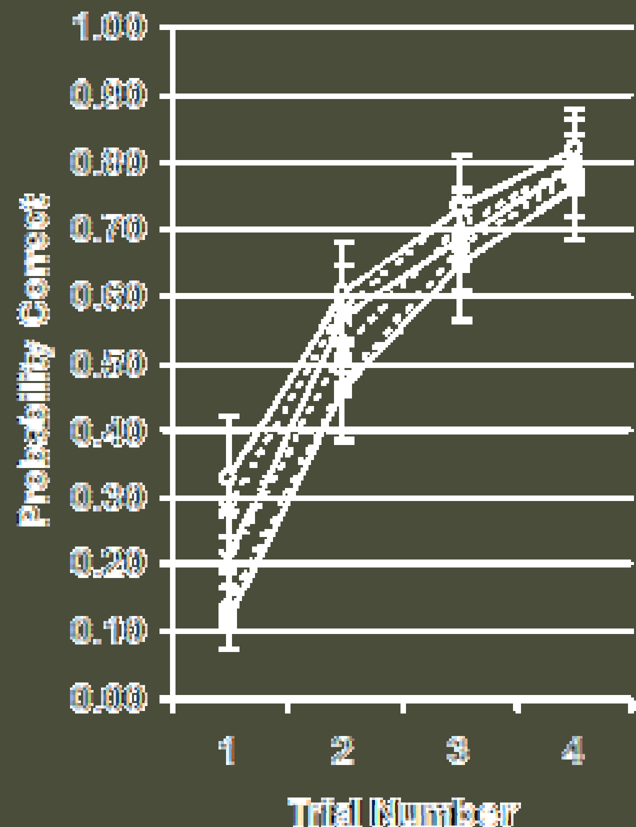
2 *SE* bars

Quantitative Results: Session 2

2 presentations (1 + 1) on session 1



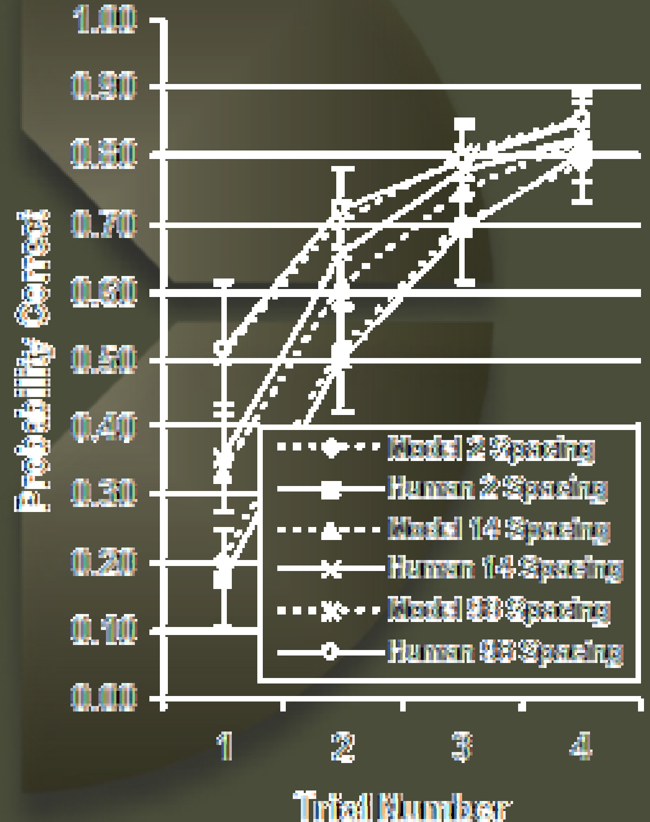
3 presentations (1 + 2) on session 1



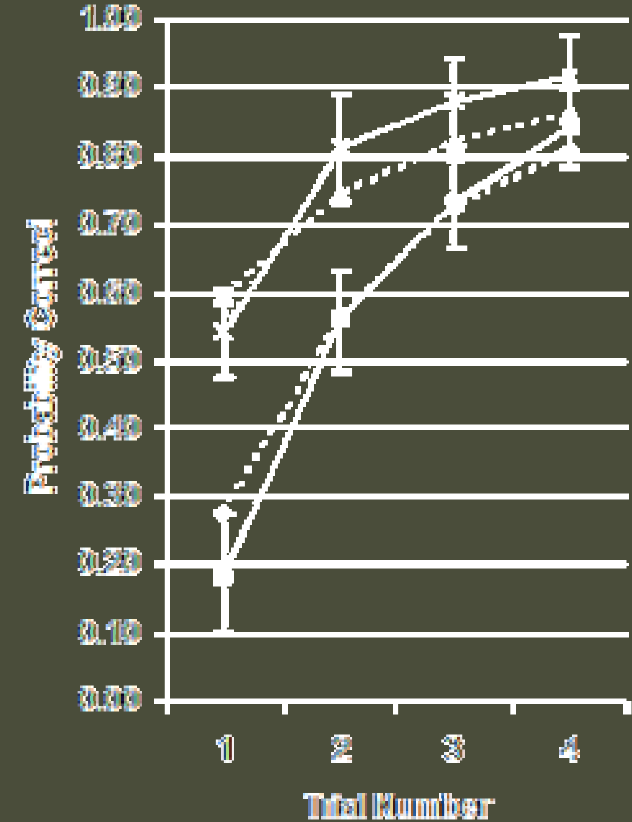
Aggregated for both retention conditions (1 and 7 days)

Quantitative Results: Session 2

5 presentations (1 + 4) on session 1



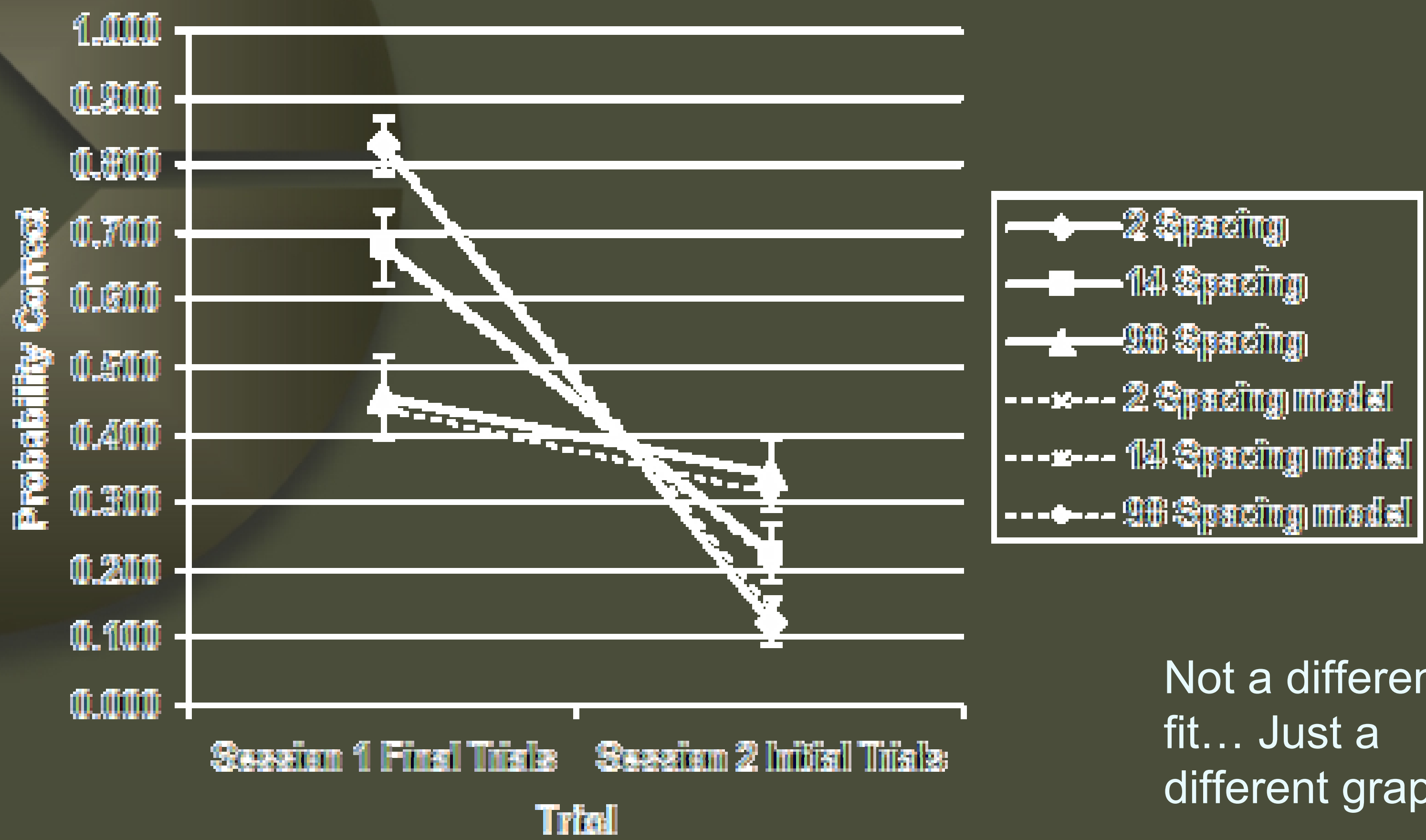
9 presentations (1 + 8) on session 1



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



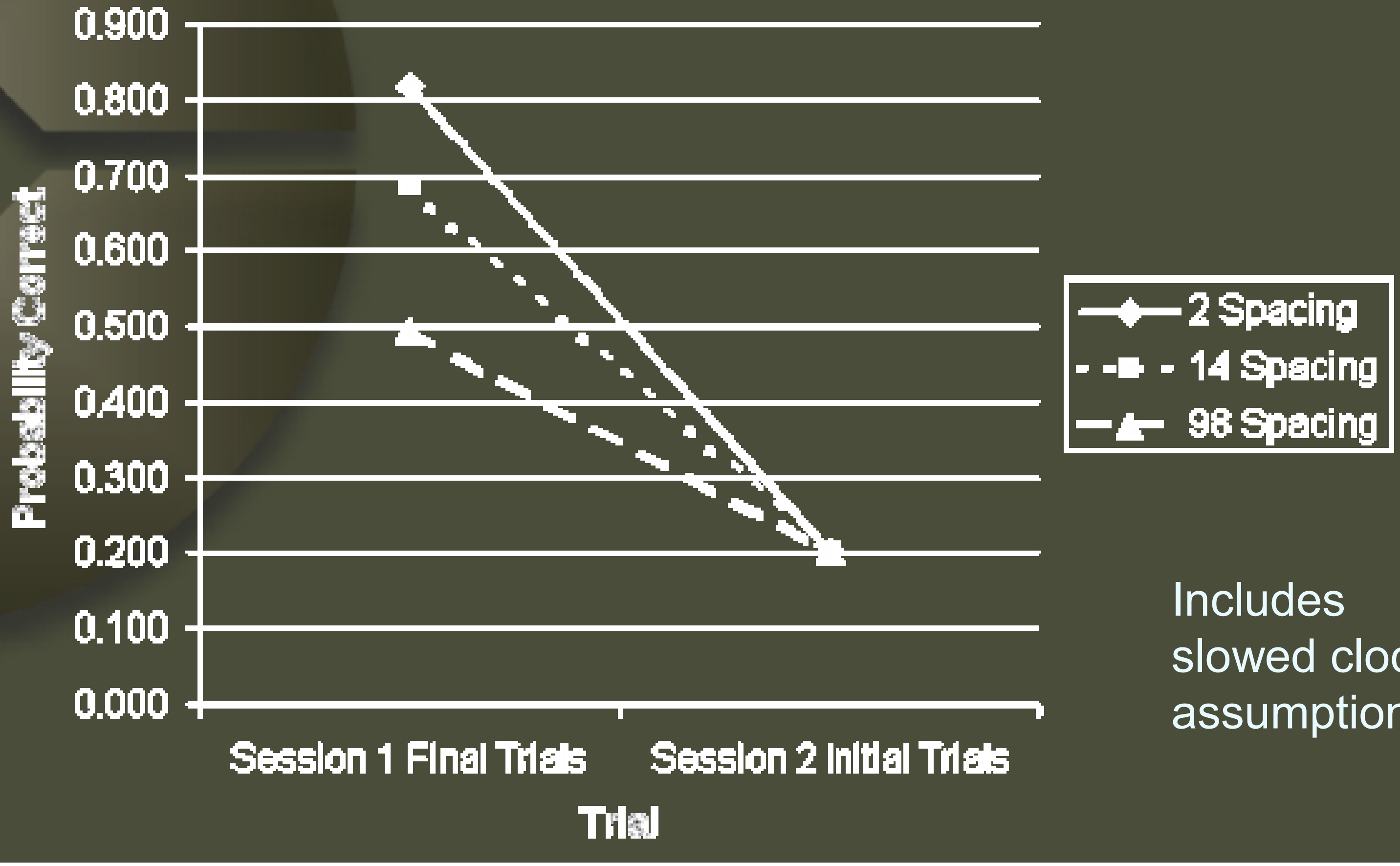
Not a different fit… Just a different graph

The Model Needed 2 New Mechanisms

- For Spacing
  - Similar mechanism to Anderson and Schooler (1991)
    - decay = f(time from previous presentation)
    - Wider spacing → less decay
  - New mechanism proposes that higher activation results in greater decay
    - decay = f(activation)
    - Wider spacing → less activation → 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

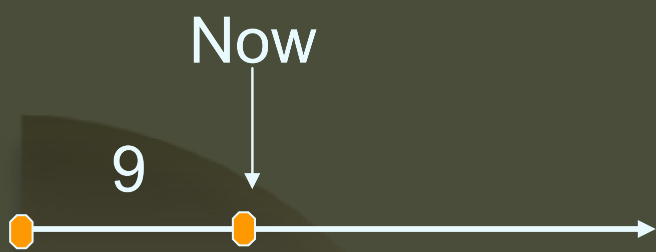


Includes slowed clock assumption

Decay Equation

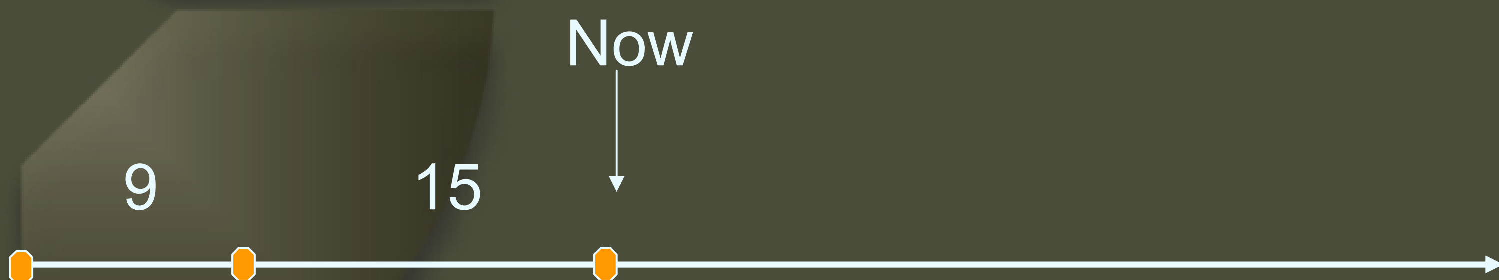
$$d_j = ce^{m_j} + a \quad \rightarrow \quad m_i = \ln\left(\sum_{j=1}^{n} t_j^{-d_j}\right)$$

- 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 → more forgetting → less activation → less decay → better retention
- Better fits than Anderson and Schooler (1991)
  - Also Anderson and Schooler (1991) tends to over-predict performance with large amounts of practice



$d1=a$ $m=\ln(9^{-d1})$

$d2=a+ce^{m}$



$m=\ln(24^{-d1}+15^{-d2})$

$d3=a+ce^{m}$

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

DEMO

Parameters and Model Statistics						
	Exp.	Bahrick (1979)	Glenberg (1980)	Rumelhart (1967)	Young (1971)	Glenberg (1976)
Parameters						
decay intercept *a*	.170	.208	.120	.143	.294	.058
decay scale *c*	.219	.138	.366	.488	.415	.276
threshold (τ)	-.679	-.679*	-.679*	-.679*	-.679*	-.679*
noise *s*	.255	.255*	.255*	.255*	.255*	.255*
h scalar *h*	.031	.031*	.031*			
encoding (b)		3.67				.373
encoding (b_r)						.289
Fit Statistics						
r^2	.944	.929	.940	.925	.461	.944
RMSD adjusted	.045	.058	.052	.021	.025	.026
χ^2	322	246	90.7	42.1	8.77	31.8
χ^2 *df*	157	27	40	38	16	20

Only decay parameters varied

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?

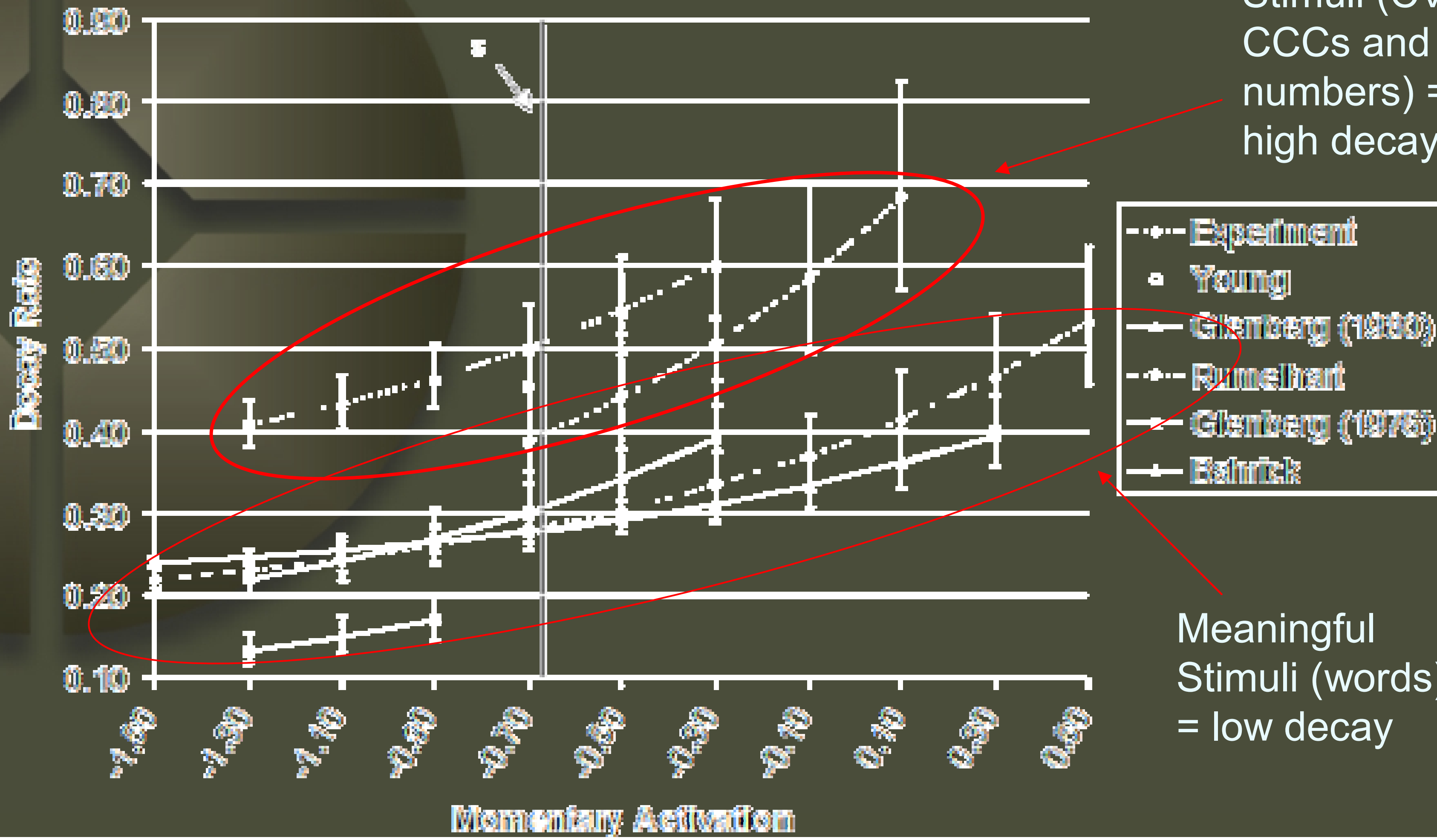
- It indicates that the level of performance and variability of performance can often be captured with only the decay parameters
- However, we do not claim that threshold is necessarily fixed

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



Meaningless Stimuli (CVCs, CCCs and numbers) = high decay

Meaningful Stimuli (words) = low decay

More specifically the slope of the lines (defined by c parameter) may represent resistance to additional practice caused by stimuli or task

	Exp.	Bahrick (1979)	Glenberg (1976)	Rumelhart (1967)	Young (1971)	Glenberg (1980)
Original Parameters						
decay intercept *a*	.170	.208	.058	.143	.294	.120
decay scale *c*	.219	.138	.276	.488	.415	.366
Forced Grouping (2 levels of c and best fitting a)						
decay intercept *a*	.181	.181	.071	.150	.288	.115
decay scale *c*	.180	.180	.180	.450	.450	.450
Fit Statistics						
χ^2 old	322	246	31.8	42.1	8.77	90.7
χ^2 new	391	328	32.6	48.2	8.80	104
χ^2 *df*	157	27	20	38	16	40

More varied associations

These experiments had meaningful stimuli and intentional learning

These experiments had meaningless stimuli or incidental learning

Contributions of the Model

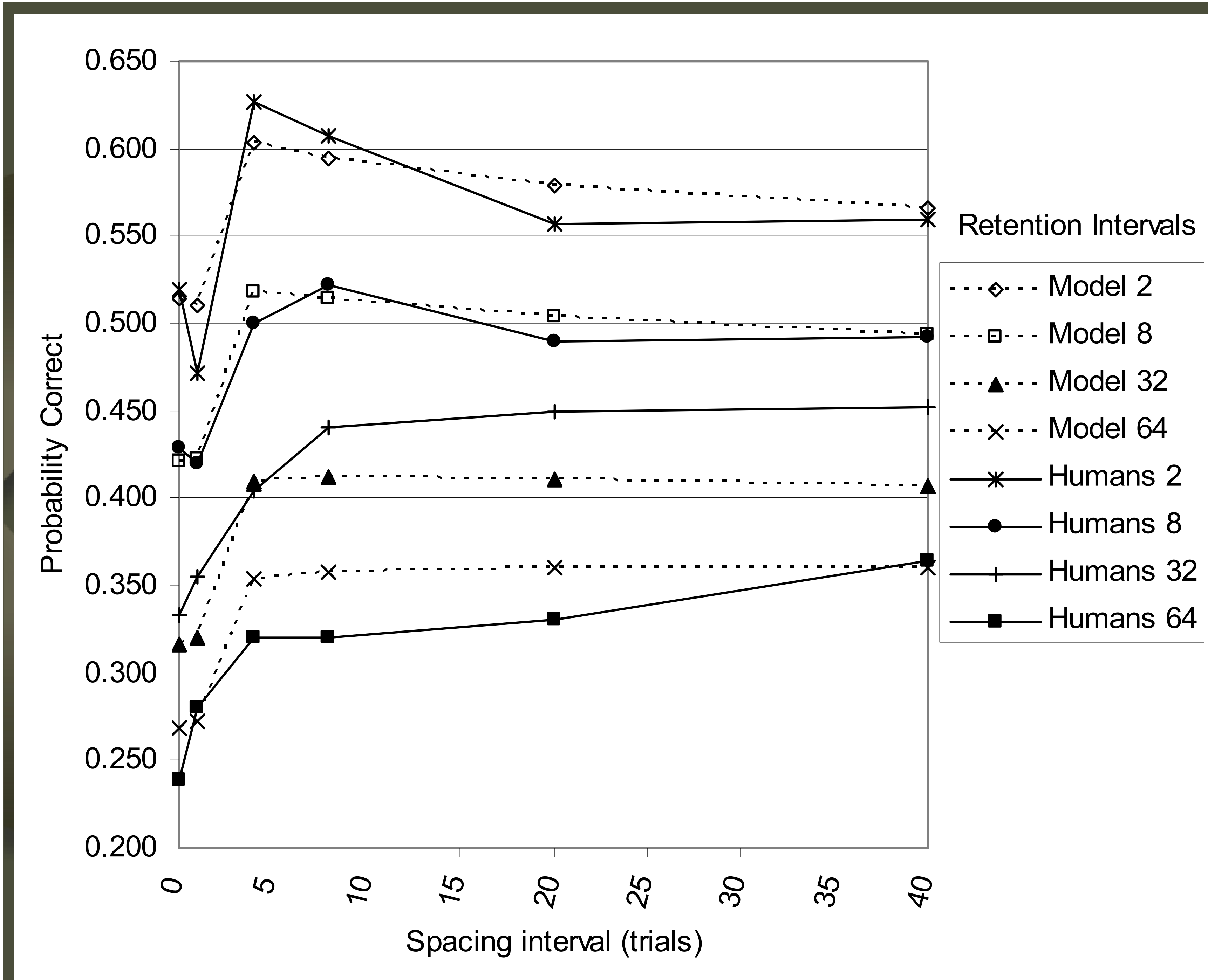
- Additional practical power of ACT-R
  - Now handles spacing well
  - Backwards compatible
    - Can be turned off by setting c=0 so using this mechanism would not be a big commitment if it was available in ACT-R
    - Implies .5 is not a bad estimate of decay
  - Since other parameters can be fixed
    - 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

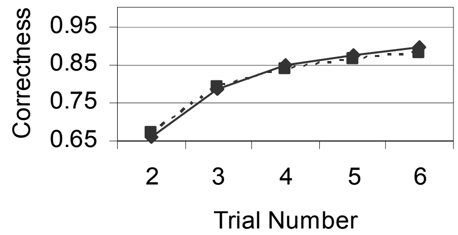
- Variable Encoding (Martin, 1968)
  - More varied contextual elements encoded for spaced trials
- Voluntary Encoding/Rehearsal (Atkinson and Shiffrin, 1968; Rundus, 1971)
  - More rehearsals/time in short-term memory for spaced trials
- Consolidation (Peterson, 1966; Landauer, 1969)
  - More time for 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?

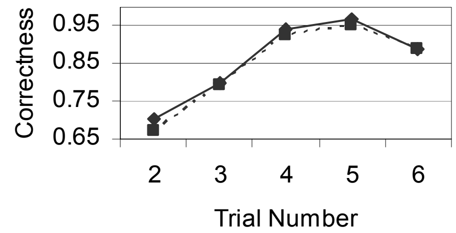
Locus of Effect	Voluntary	Involuntary
Between P1-P2	Rundus (1971) *Rehearsal*	Landauer (1969), Atkinson and Shiffrin (1968) *Consolidation or STS transfer*
At P2	Shaughnessey et al., 1972) *Attention*	Mensink and Raaijmakers (1988), Hintzman (1974), Bjork (1969) *Contextual Fluctuation, Habituation, or Effort*
After P2	Greeno (1970) *Decision to Forget*	Our Model Wickelgren (1973), Anderson and Schooler (1991) *Forgetting*



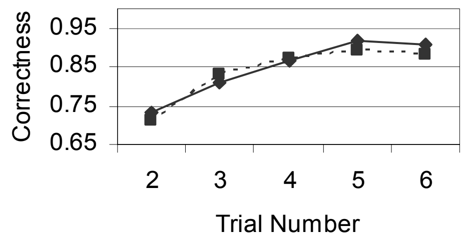
0.650
0.600
0.550
0.500
0.450
0.400
0.350
0.300
0.250
0.200
Probability Correct
0
5
10
15
20
25
30
35
40
Spacing interval (trials)
Retention Intervals
Model 2
Model 8
Model 32
Model 64
Humans 2
Humans 8
Humans 32
Humans 64



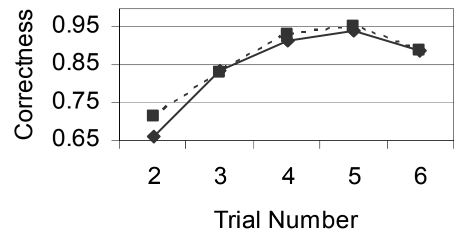
(a) 10-10-10-10-10 Spacing



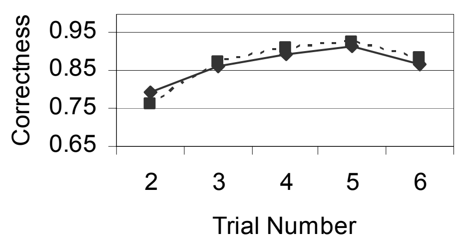
(b) 10-10-1-1-10 Spacing



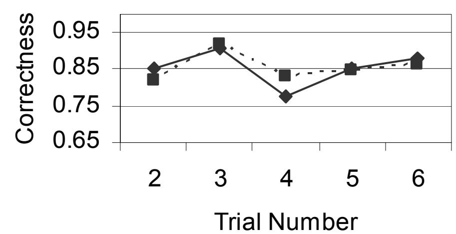
(c) 6-6-6-6-10 Spacing



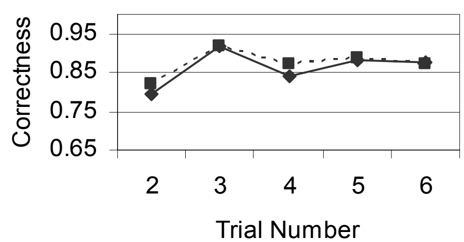
(d) 6-6-1-1-10 Spacing



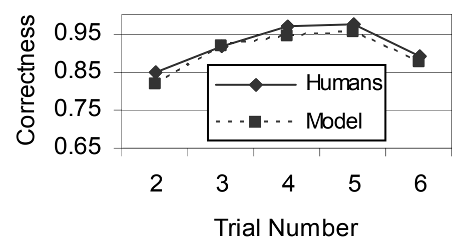
(e) 3-3-3-3-10 Spacing



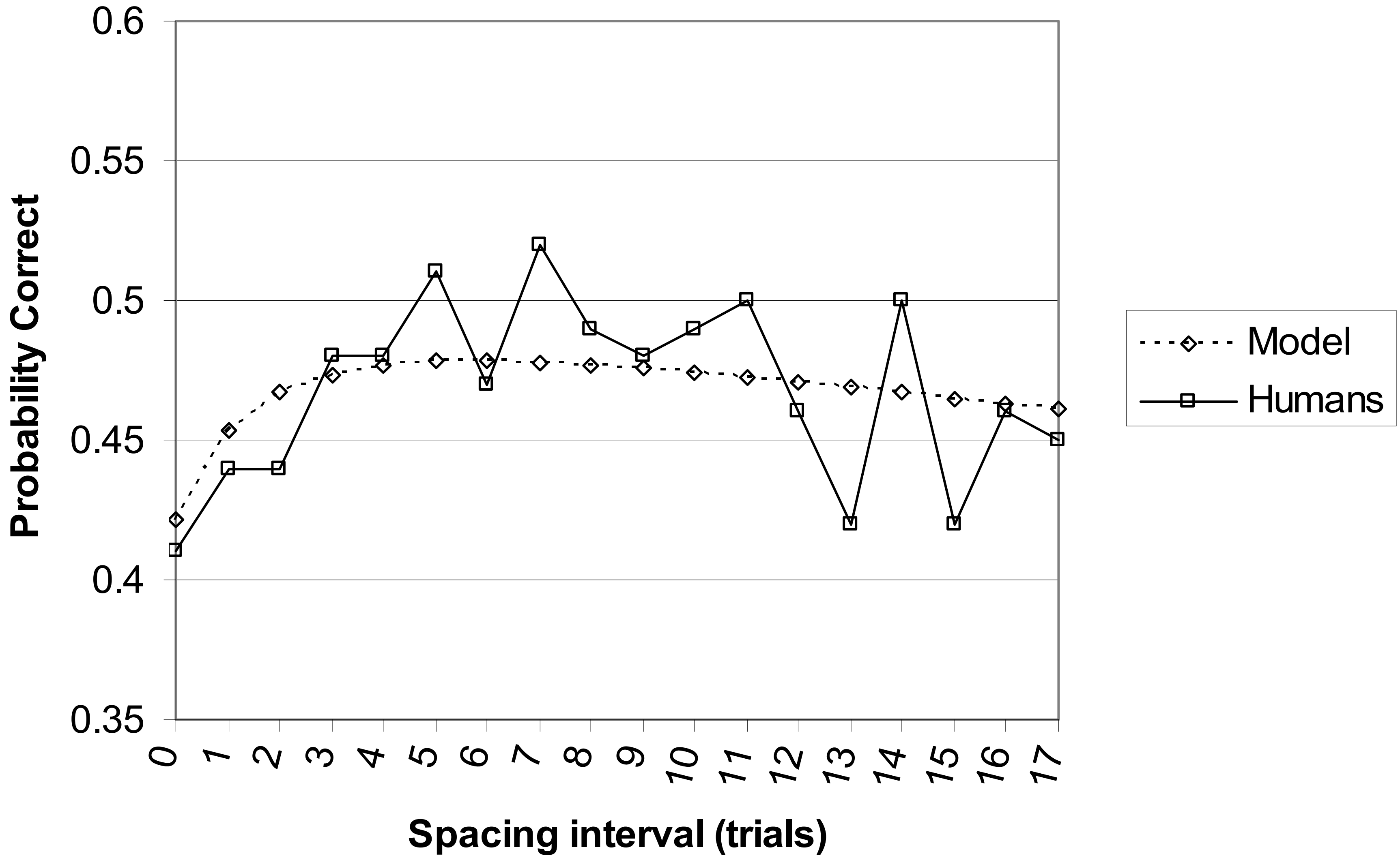
(f) 1-1-10-10-10 Spacing



(g) 1-1-6-6-10 Spacing



(h) 1-1-1-1-10 Spacing



Probability Correct
0.6
0.55
0.5
0.45
0.4
0.35
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
Spacing interval (trials)
Model
Humans