Lea:

I will assume the new study has 50% target tests, and after list one four types of foils: A) Novel; B) A foil on trial n-1; C) an item studied but not tested on n-1; D) an item studied and tested on n-1.

On list one the result look normal: Good performance and CRs a little higher than Hs.

On list two (now with four types of foils) H’s drop about 5%. A CRs stays about the same. CR’s for B, C, and D are terrible, from about .5 to .6. I assume that during list two testing they are learning that they are making many errors on foils (I asked you to look at test position effects for the first three lists to see if we can see learning taking place. This depends on when and what type of feedback is provided).

On list three they seem to adapt from what happened on list 2: The main change is a further decrease in overall performance, most notably for targets: Hs: .79 (list 1); .72 (list 2); .58 (list3). Only foil type A (novel) remains good (.76).

Why the large drop in H’s from list 2 to 3? Assume that the match of the target to its trace, call it MT, is the same for list 1, 2, 3. The likelihood ratio for the target trace (call it TL) cannot remain constant for the three lists, not can the odds O used for decision. Suppose the odds O used for a target decision is the same for lists 1, 2, 3. Then the only way to get a drop in H would be large shift in decision threshold, but that would INCREASE CRs, but the CRs in list 3 if anything decrease. Thus it seems that O and TL must have dropped on list 3.

O is p(Old|D)/p(NEW|D). REM assumes the denominator is based on random matching of a foil. But most of the foils on lists 2 and 3 are now highly similar to the list traces (on list n-1), due to what we presume is a difficulty using context to focus only on the current list. It is possible that both the subject S and REM calculate and use a corrected O based on an adjusted denominator, assuming a mixture of one random foil (A), two highly similar foils (B and C), and one even more similar foil (D). The distribution of O for a target test would then (I think) shrink toward 1.0, even if MT remained constant.

O for foil tests and target tests would now overlap a lot, but the type of foil tested would matter: Type A would not have a strong match on any pf the current of prior lists. Type B, C, and D would have a string match. I think this model could produce the observed results, roughly, for list 3, at least if there were little ability to focus only on the current list.

What about list 2? List 2 could be less extreme because the highly similar items are only from list 1 (n-1) and not from list n-2. In addition, the S will not know the foils are highly similar at the start of list 2, and some learning will take place during the testing (and that depends on the feedback – so let me know the exact way feedback occurred).

A separate idea worth considering is rather different: Faced with similar foils caused by propr list confusions, S could focus more on current list context (changing context) hoping to stop activation of similar items from prior lists. This could occur at storage, at test, or both. The gradual increase in performance as lists continue after list 3 suggest something like this is going on. However if does not seem to have worked well (or at all) for list 3. But suppose S has to learn how to do this and at list 3 does not know how, but nonetheless diverts content storage to context storage. If the context used was not yet helpful for discriminating lists, this would harm both target and foil tests without other changes in REM (such as those discussed earlier): It would be as if u and c parameter values dropped.

At the moment I don’t have a good idea concerning what combination of the two ideas just mentioned might be operating. Perhaps the final testing results will provide some clues.