Chapter 6

Discussion and Recommendations for Further Research

Summary of Hypothesis Tests: By and large, our five hypotheses were not supported. Our first and second hypotheses have some visual support (i.e., recall proportions across sentence types start out roughly equal for low PE subjects then branch out; high PE subjects seem to treat sentence types differently from the start); however, these differences were not statistically significant.

Analysis of our third and fourth hypotheses was confounded because our candidate contaminating covariate failed to have consistent effects. This, coupled with the floor effect of our PE scale, the unexplained (and substantial) variability in recall behavior, and some other control issues (detailed below), made the use of our continuous DVs less than fruitful. The floor effect of the PE scale was especially problematic – with many subjects compressed at this floor, relations would be difficult to see even if present. In an attempt to detect weak effects of prejudice, we aggregated subjects by PE (as in high and low prejudice). Aggregation probably made the floor effect-driven range restriction less problematic (the subjects lumped together on PE's floor are probably less-afflicted with well-practiced prejudicial expectations than the high half of PE scorers). This exercise generated weak support for our third hypothesis: the time interval data feebly indicates that high PE subjects manifest a negative impression-centered person-memory schema in their storage of sentences about a Black target – and, unlike the low PE subjects, they apparently do this starting with the earliest blocks of sentences.

The median split approach failed to generate support for our fourth hypothesis – where we expected to see bolstering replace inconsistency resolution (in the slow condition) since subjects were afforded the time. There was weak evidence, however, that more inconsistency resolution was occurring in the fast condition. This evidence was in the form of greater recall time interval differences seen when comparing high PE subjects and their schema-speeded versus non-speeded intervals. The bottom line for the first four hypotheses is still this: we failed to create a condition where prejudice would paradoxically favor recall of laudable or admirable inconsistencies associated with a fictitious Black target. We cannot challenge Stangor and Ruble's (1989) assertion (see discussion below).

Our fifth hypothesis was just intended to verify that racial prejudice does not predict recall behavior when the target is White and so are the subjects. So using a White target, we performed the same sort of tests seen above. Fortunately, relations with PE ranged from weak to very weak – and, of course, were non-significant.

In sum, these outcomes suggest that H-SAN processing *effects* may not reliably manifest themselves in the prejudiced rater/performance appraisal arena — at least not in designs similar to those used previously to illustrate H-SAN effects. In the previous chapter, we also indicated our *reluctance* to suggest that H-SAN memory *processes* do not generalize here. We believe our reluctance is merited by consideration of the following:

Stimulus Presentation Speed: Our comments in this final chapter are largely driven by the pitfalls or disappointments experienced in executing the study just described. The greatest of these all relate to our inability to counter Devine's (1989) conclusion that racism scores become less important as cognitive processes approach automaticity and Stangor and Ruble's (1989) conclusion that prejudices will not favor recall of incongruencies even when information rates are high. The problem here relates to our earlier insinuations about stimulus presentation *method effects*. In other words, fully automatic sentence

presentations on computer monitors may allow greater subject concentration than timer-prompted page-turning of sentences on separate sheets of a three-ring binder. This, coupled with our software's unfortunate expansion of the 8 and 12 second rates by a full second, may have pushed us out of the realm of dominant inconsistency resolution.

In our background chapter, we suggested that use of bolstering opportunities — provided in small pieces (i.e., during on-line processing) — might be more *habit*-driven than post-stimulus bolstering in a single lengthy (i.e., relatively unconstrained) block. We, therefore, opted for offering on-line bolstering opportunities alone. Obviously, our desire was to catch the manifestations of such prejudice-driven *habits*. Unfortunately, 13 seconds in a video format may have been too long (inducing boredom and aggravating fatigue effects) while 9 seconds may have been *long enough* to allow fast readers time for on-line bolstering. This may underlie the positive correlation ($r = .52$, $p \leq .001$, $n = 32$ subjects providing 608 recalls) between PE and Congruency Recall Advantage in the 9 second/Black target cell of our data set. Or it may be that Stangor and Ruble are correct in suggesting well-entrenched expectancies are unlikely to lead to enhanced recall of incongruencies under any conditions.

These possibilities lead to our most straight-forward recommendation — adding cells to our 13 and 9-second design to study a four-second deviation in the opposite direction. We would add a White cell and a Black cell that each use 5 seconds of exposure per stimulus sentence. This should effectively eliminate any opportunities to bolster congruencies on-line. To allow straight-forward comparisons to the 9 and 13-second cells, we would make no changes to experiment materials or procedures (like those below) before collecting this data. We would also add a White target cell at the 9 second level to verify that White-subject PE operates only when the target is Black. And that would finish our use of this study's existing materials — forever (though we should note: modification of these materials to study H-SAN effects in an Air Force leadership potential/performance appraisal/promotion board context is underway).

Sequence of Judgment versus Recall: It is at least possible that our results appear more supportive of Stangor and Ruble's (1989) position — versus that of Wyer, et al (1987) — because the former applied the surprise recall test *before* the judgment form — just as we did. We used such a sequence because we were concerned that the act of finalizing and reporting overall assessments of the target would alter the target's representation in memory. Specifically, some pathways might be strengthened — or even built anew. We cannot rule out, however, that the sequence we used made it difficult to replicate the Wyer, et al (1989) findings in a prejudiced rater/performance appraisal context. Future research is still needed to compare what differences there may be when this sequence is altered (again — within a prejudice/performance potential domain).

Irrelevant Versus Neutral Items: Our literature review allowed us to make predictions concerning the diminished recall of neutral/irrelevant items. In H-SAN related research, these are expected to be recalled least of all. Such predictions are related to conclusions of previous researchers — who've: "postulated that irrelevant /neutral items are likely to have connections only with the person-concept [or super-ordinate] node. This is because irrelevant/neutral items have no other logical connection to the other behavioral items except by association with the person performing the behaviors" (p. 20, Chapter 2).

Although our data marginally support the lower recall proportions predicted for such items (P_{ALL} = .337, P_{iR} = .282, $p < .01$, N = 79 subjects providing 1,566 total recalls), the inter-recall time intervals suggest some direct inter-item linkages between these neutral items and incongruencies. We suspect a competing complexity not previously considered. We unfortunately ignored the notion that full-time students are always "on-duty" as it were. In other words, most if not all descriptions of Jermaine's supposedly irrelevant activities could be viewed as actions pursued in place of study or homework. This makes them not so much irrelevant as simply neutral. As such, they could have been treated as neutral congruencies (as in *departures from academics*). This would explain the long inter-recall time intervals between iR's and C's and the very short intervals between iR's and I's (see Figure 18, p. 78).

What remains a mystery is the short intervals between successive iR recalls (again, see Figure 18). We suspect our subjects compared these supposedly irrelevant neutrals to each other to minimize their confusion. Such activity would build links between them without necessarily building links between them and intended congruencies.

To avoid such complexities, we would recommend that future studies of this sort have candidate stimulus sentences rated for pertinence — not just negativity or neutrality as ours were. Though we wrote an extra 21 candidate sentences (so we could throw out those that were ambiguous in terms of our three categories: congruent, neutral, or incongruent with a negative stereotype), we'd recommend far more candidates in future work. This might allow identification of sentences that are not just neutral — but *truly* irrelevant. Such should allow a more direct test of H-SAN effects in this arena. We should emphasize *this arena*, since our tasking requested subjects to assess the target's worthiness for continuation in a rather generous assistantship. Anything viewed as even marginally pertinent might be treated differently than neutral/irrelevancies were treated in prior H-SAN research (where assessments usually relate to affability or friendliness).

Interaction Between nCog and Believability: Related to the above problem is the notion of believability. The reader may recall the following from our background chapter: "low nCog subjects may only apply critical consideration to information from dubious sources. [More to the point]....subjects low in nCog may only think critically about information from sources they view as.....suspect" (p. 23, Chapter 2). We've treated inter-item comparisons as a component of such critical thinking, and we've predicted that inter-item links result from such comparisons. However, we failed to assess our stimulus sentences for *believability*. As with the previous problem (above), we'd recommend future designs in this arena assess sentences along these additional dimensions before finalizing the stimulus set.

Greater Consideration of Workload Perceptions: The enormous magnitude of our subjects' assigned information processing task probably does not *sink-in* until the third or fourth or fifth block of sentences (depending on their reading speed and/or their personal tolerance for verbal processing *effort*). Once it does *sink-in*, it might lead our subjects to consider dropping the optional bolstering activities and instead pursue arguably less-optional *inconsistency resolution* efforts.

Since the only strong evidence we found for the latter was the comparatively longer intervals between successive congruent recalls, we're tempted to dismiss this potential confound. However, future studies in this area might consider ways to stabilize chosen processing strategies.

Pre-experiment practice sessions using a different target ratee might give subjects a preview of the upcoming work-load. A better test of H-SAN effects (given an entrenched or well-practiced expectancy) might be in a design where subjects choose a processing strategy early-on based on information rates — then stick with it. Otherwise, it is difficult to split apart changes in recall proportions due to *impression formation* versus *work-load frustration*.

High PE-driven Self-consciousness and Hypothesis Guessing: Future trials should collect hypothesis-guessing data to see if high PE scorers are more aware of the study's purpose. This could distort proportions in obvious ways. Distortion of inter-recall intervals would likely be due to pauses (where hostile or offended or self-conscious subjects might deliberate seeking only incongruent memories). This problem may be related to the wide variety of standard deviations observed across our six cells. Though this did not prevent paired-sample (i.e., correlated sample) t-tests *within* each cell (for Hypothesis 3), it did make us more skeptical in our comparisons of these t-test results across cells. For instance, in one case a five-second advantage for schema-speeded inter-recall intervals was significant. In another cell, it was not.

Conveniently, this produced a pattern of results supporting our hypotheses (where little else did). It still requires a leap of faith to believe positively that the fortuitously tighter variance in some of these cells was driven by our proposed causal processes. We would say *probably* (or should we say *hopefully*). Future studies of this sort should declare *a priori* a set of criteria for discarding peculiar time intervals. If similar results are again obtained, theoretical consideration should be directed at predicting the variety of dispersion cell by cell (and not just the expected time interval differences).

Confounding by Impression Management: We just mentioned that hostility in high PE subjects might (via self-consciousness) distort recall behavior. The problem may really be larger than that and should possibly be couched in terms of impression management. In at least two ways, IM might be working directly to distort observed relations between PE and recall behavior. The obvious confounding mechanism relates to the possible reluctance of self-conscious subjects to report their congruent recalls (alluded to above). If these subjects were also at greater risk of under-rating their PE-related attitudes, this might have exaggerated the predicted PE to CRA correlation in the slow condition (had it not been for the unfortunate floor effect of the PE scale).

The second distorting mechanism is again related to the above – but is possibly more pervasive and less obvious. As we said, IM in a self-conscious subject could make them especially reluctant to report a congruent recall after they've just reported one or more previous congruencies. This is especially problematic in designs like ours where congruencies dominate the sentence list. The time intervals between successive congruent recalls are likely to be expanded by a self-conscious subject's stubborn attempts to continue searching memory of non-congruencies (i.e., stereotype inconsistencies or irrelevancies). This could lengthen all intervals beginning with a consistency – but would probably lengthen

the C to C intervals most. Of course, we expect lengthy C to C intervals when anticipating H-SAN structure, so this alternative explanation is disconcerting. We need to deliberate further about the differential effects of IM on our self-report measures versus our experimental task.

Hauenstein (personal communication, 1995) thought some consideration should be given to use of a bogus pipeline. In our design, we might have wired subjects to the computer in front of them and implied that obvious deceptive practices during the self-report phase would disqualify them. There are two reasons we did not choose this course in our study. First, my USAF-assigned mission here was to develop an expertise in psychological measurement. Therefore, the emphasis was always on development and validation of the PE scale. Since a bogus pipeline had not been used with the initial scale development sample (N=397), I was reluctant to use it in the lab. This particular concern was overcome by events (and by results). As is apparent from comparing Figures 3 and 4 in our third chapter, we ended up with two somewhat distinct populations anyway (probably a "we feel fairly anonymous" population in the first case and a "we feel not so anonymous" population in the latter).

When we say overcome by events, we're alluding to our second reason for not applying a bogus pipeline. This university's Human Subjects Committee (HSC) had months worth of serious concerns about trauma (via emotional arousal and invasion of privacy in our study). We're fairly confident they'd have been seriously troubled by the addition of bogus wiring while measuring racial prejudice. However, it might be of value someday to see what effect a bogus pipeline would have on PE scores – hopefully IM distortions would be diminished (and likewise the floor effect). Of course, it's not clear what effect this would have on subject performance in the experimental task. Perhaps increasing perceived anonymity and perceived time pressure would help for the task.

At this point, we are tempted to give scale improvement another try. Though we wrote dozens of original items – most intended to be quite subtle by comparison to those on existing scales – these did not survive the multiple hurdles outlined in Chapter Three. Typically, they failed to load cleanly or unidimensionally on the factor we called PE. Our recommendation for future efforts here would focus on increasing the subtlety of the eleven surviving items. Hopefully, the degree to which they evoke IM can be diminished.

Fatigue: The above-mentioned confound with impression management could also have been exacerbated by fatigue effects. At the beginning of this chapter, we mentioned our concern about lengthy intervals coming late in our subjects' recall performances. After answering nearly two hundred questions on a survey, then reading several pages of instructions followed by 56 sentences followed by a U. S. states and capitals distractor task, these folks were probably getting tired (just reading the last three sentences was tiresome). It is reasonable to expect cognitive processing speed to be slowed by the end of several lengthy mental tasks performed in series. Fatigue could have affected attention (and processing efficacy in general) for the later parts of the stimulus set. It may have likewise hampered recall performance in the last minutes of the experiment. Fatigue could also have made the cognitive load more salient. Cognitive load has been seen to differentially affect recall for stereotype-consistent versus inconsistent information (McCrae, 1997). The differential effect may depend on the degree to which subjects are dedicated to impression management. In other words, high IM subjects may spend so much effort trying to suppress stereotype-consistent thinking, they fail to deeply process the inconsistencies.

Fatigue could also have lengthened the recall intervals – especially the later ones. If subjects – possible motivated by impression management – postponed recall of consistencies until they could remember nothing else, we'd not only see such intervals preceding consistencies lengthened (as mentioned above), but we'd also see more of these intervals in the later recalls where fatigue might be more prevalent. We might suffer this in ways that would not affect prior H-SAN experiments that use artificial (i.e., lab-induced) expectancies. For these earlier efforts, there was likely little disgrace or stigma associated with recalling negative sentences. There might still be simple fatigue effects in early H-SAN work, but probably not the potential confound with impression management (and hence PE) that we're forced to suspect in our work. This fatigue influence would probably operate on top of the previously-mentioned delays due to re-visiting prior recalls. It would probably make such re-visits even more frustrating.

Intrigue-driven Memorability: We mentioned above the need to consider both believability and stereotype-consistency when examining the apparent probability of a sentence's later recollection. It might also be appropriate to have subject-similar judges rate these sentences for their memorability – especially in terms of entertainment value or intrigue. Some of the stimulus sentences were more likely than others to catch and hold our subject's attention (e.g., "When Jermaine Washington couldn't find a smaller guy's date to flirt with, he'd spend the night before a test playing basketball"). Though such an intrigue-driven analysis could be designed into future studies, it could also be applied to the existing data set.

Priming Effects of the Survey: Though we designed our study to prime a presumably existing anti-Black prejudicial expectancy only in subjects from my treatment cells, we may have primed all subjects before the computer-based exercise even started. The on-screen introduction for the treatment groups described an African-American student from an economically-disadvantaged background. However, this was not the only source of priming. It would be reasonable to expect many of our subjects to have several working hypotheses (about the purpose of our study) as they worked through the pre-exercise survey. These were likely to include racism-related efforts (though the bulk of the survey measured other phenomena). Since race-related attitudes where tapped by some items, racial prejudice-related expectancies were likely primed (especially negative ones given items like "To what extent do you agree or disagree: African-Americans would rather accept welfare than work for a living"). We alluded to this earlier when discussing the strange recall patterns seen in the White-target cell. We mention it here to suggest the possibility that survey-induced priming may have diminished the impact of pre-existing levels of PE on our recall data (i.e., it's troubling when using PE to identify subjects afflicted with highly accessible negative expectancies when you've primed the expectancies in everyone – doubly so when it confounds the treatment).

Final Thoughts: This dissertation effort was to be the preliminary foundation for an ambitious research program. The program's ultimate goal was the development of new validation methodologies for Air Force human relations courses — especially the initial six to eight hours of coursework received by Air Force recruits. Our results, so far, are far too weak to suggest this sort-of application is near-at-hand. We hope that by pursuing the above recommended improvements, we'll progress towards such

an end. Fortunately, our lab subjects' task (reading 56 sentences about a stranger to form an impression of his performance potential) is not terribly dissimilar from the work of Air Force promotion boards. So, at the very least, the software programs and laboratory protocols we've developed for this dissertation should have some future value for the Air Force. Non-traditional USAF minority groups claiming prejudicial mistreatment by these boards include navigators, female officers, prior-enlisted officers and members of non-flying operational career fields (e.g., missileers). Adding in racial minorities, potential USAF applications are numerous. The Air Force would hope to find that, even if highly prejudiced board members treat sentences differently than their colleagues (a big if), these members' final judgments — and the validity of the final rankings — do not suffer. We've not even encoded our own judgment data yet — so we won't speculate.