The Temporal Dynamics of Word Comprehension and Response Selection: Computational and Behavioral Studies

Blair C. Armstrong

August 10, 2012

Psychology Department
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:

David Plaut, Chair
Marlene Behrmann
Brian MacWhinney
Natasha Tokowicz

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2012 Blair C. Armstrong.
Distributed under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 License.

Funding for this research was provided by Alexander Graham Bell Canada Graduate Scholarships from the Natural Sciences and Engineering Research Council of Canada, by Carnegie Mellon University, and by the Center for the Neural Basis of Cognition.

Abstract

Most words are semantically ambiguous in that their meaning depends on the context in which they occur (e.g., <river> vs. <money> BANK). Developing a theory of how the meanings of semantically ambiguous words are comprehended has proven difficult because of discrepancies in the effects of relatedness of meaning observed across tasks. Further, existing accounts are underspecified, narrow in the scope of issues they address, and mutually inconsistent.

The current work proposes a theory of semantic ambiguity resolution in which the discrepant effects are explained by the temporal settling dynamics in semantics and how these dynamics interact with the semantic representations of ambiguous words. This account was instantiated using a distributed connectionist network that incorporated biologically processing constraints. The network shows that the semantic activity evoked at different points in time is consistent with the effects observed in different tasks. The account is further supported by behavioral studies of lexical decision, to evaluate whether differences in processing time, as opposed to qualitative task differences, are responsible for the different ambiguity effects observed across tasks. In these experiments, task difficulty—and the presumed amount of semantic processing—was manipulated both by altering the wordlikeness of the nonword foils and by altering the visual contrast of the stimuli. The selection of optimized word stimuli was enhanced by the development of an automatic stimulus selection algorithm which allowed for a large number of confounding variables to be controlled for, including a measure of the relative frequency of an ambiguous word's meanings that was collected using a new norming method. The results of the lexical decision experiment show that the contrast manipulation caused large increases in overall latencies and produced semantic ambiguity effects consistent with later semantic processing. This coordinated computational and behavioral work suggests that properties of settling dynamics within a distributed network explain the discrepancies observed across tasks, and generate predictions that can guide future research.

Furthermore, this work points to the importance of understanding how the semantic, orthographic, and phonological representations interact with the response selection system to generate the patterns of effects observed in different tasks. The second portion of the dissertation develops a model of response selection that employs a similar set of domain-general learning, processing and representation principles to those that were used to model semantic ambiguity effects. This work was challenged by previous computational and behavioral investigations using a numerosity judgement task that revealed numerous disagreements between the connectionist models and the behavioral data. New behavioral data collected in an extension of the original numerosity judgement paradigm show that some of these findings do not replicate and were likely due to several idiosyncratic aspects of the original experiment. Connectionist simulations of this extension of the original task succeed in capturing key elements of these new data, including some that are not captured by other models. This work provides the foundation for developing models that integrate the word comprehension system and the response selection system to understand and predict the effects of ambiguity in different tasks, and beyond.

keywords: word comprehension; semantic ambiguity; temporal processing dynamics; models of response selection; computational/connectionist modeling; lexical decision; stimulus degradation; numerosity judgment; selecting optimized stimuli; norming meaning dominance

Acknowledgments

Over the course of my graduate training I have had the priviledge of interacting with a number of individuals who have guided and improved my work in innumerable ways. My advisor, Dave Plaut, has been a central figure in all aspects of my development as an independent researcher. Dave has always been a fountain of knowledge no matter what issue has challenged me in the course of my work. Intertwined with his sharing of this knowledge, however, has been the even more valuable opportunity to learn how to think about the issues within a given domain. From this experience, I have developed a deep appreciation (and hopefully, a fragment of his capacity) for identifying and critically thinking about "where the action is at". I am also grateful for how he has encouraged me to push myself in developing my own expertise—being willing to hash out an issue when I was stuck but giving me the freedom to learn how to learn independently as well. This has allowed me to discover what I know, what I do not know, and how to move from the latter to the former. For all of this and more, thanks Dave.

I am also very grateful for having an outstanding thesis committee throughout my graduate training. Marlene Behrmann has been there since day one of my training and has been a wonderful source of insight, enthusiasm, and feedback that has helped me discover both the strengths and weaknesses in my work and how to improve them. Brian MacWhinney has always asked great questions that probe my theories in critical ways that I often did not realize until I found myself being asked the same questions in various other settings. Natasha Tokowicz has been extremely generous both with her own expertise and in opening up her lab and the reading and language community at the University of Pittsburgh to me—I have learned immensely from those additional interactions and I could not have accomplished all that I have without her help.

I would also like to thank Christine Watson and Sarah Laszlo for their help and support over the past several years. I cannot imagine smarter or more generous lab mates and friends. You have been there in the trenches with me and shared both in the successes and challenges that are unique to connectionist modeling (which involves units AND connections) and I could not have asked for better people to do so with. Similarly, I would like to thank Tamar Degani, Alba Tuninetti, Chelea Eddington, and Alison Philips for extended discussions about all things psycholinguistic, and for their help with the practical aspects of running my experiments at the University of Pittsburgh. Beyond the immediate labs that I have worked in, I have also found a number of great colleagues and friends in Dussy Yermolayeva, Alan Jern, and Bronwyn Woods. Dussy in particular is deserving of thanks for her help and support on a number of projects within and outside my thesis work.

Special thanks are also due to George Cree, Steve Joordens, Gerald Cupchik, Doug Bors, Ted Petit, and Lucien Mantha for giving me the foundation upon which my training with Dave was based, and for their continued support as friends and as collaborators on a number of projects.

Of course, none of this work would have been possible without the backdrop that my family has provided. Thanks to my mother for fostering my initial interest in science, to my father for my interest in various non-scientific pursuits, to my grandfather for instilling in me a fraction of his focus and positive outlook (as well as for giving me hands on experience in discovering other jobs that I would not want to do for a living), to my grandmother for sustinance, and to my brother for the occasional break from it all. I am also very appreciative of Riley, Quinte, Simba, Josie, and Cece for their unconditional support.

Contents

1 Introduction and Overview **1**

I Towards a General Theory of Semantic Ambiguity Resolution **9**

2 Introduction and Review of the Semantic Ambiguity Literature **11**
2.1 Review of Key Results from the Semantic Ambiguity Literature 15
Studies that do not involve context 17
Summary 32
Studies that involve context 32
Summary 42
2.2 Theoretical Accounts of Semantic Ambiguity Effects 43
Orthographic accounts 44
Semantic accounts 45
Decision system accounts 48
Critical Evaluation of the Accounts and Objectives for an Improved Account 49

3 The Settling Dynamics Account **51**
3.1 Introduction to the Settling Dynamics Account 51
Early context-free processing 53
Late context-sensitive processing 57
3.2 The Settling Dynamics Account and the Existing Literature 58
Studies that do not involve context 58
Studies that involve context 60
Summary 61
3.3 Challenges for the Settling Dynamics Account 61

4 Substantiating the Settling Dynamics Account: Connectionist Simulations **65**
4.1 Settling Dynamics Simulation 1 66
Methods 67
Network architecture 67
Training patterns 67
Training 69

Results . . . 69
Discussion . . . 69
4.2 Settling Dynamics Simulation 2 . . . 72
Methods . . . 73
Network architecture . . . 73
Training patterns . . . 75
Training . . . 77
Results . . . 78
Settling dynamics . . . 78
Underlying mechanics . . . 78
Discussion . . . 84
4.3 General Discussion . . . 85
Future simulation work: Improved context representations . . . 86

5 Evaluating the Settling Dynamics Account: Behavioral Investigations 91
5.1 Lexical Decision Experiment 1 . . . 94
Methods . . . 94
Participants . . . 94
Apparatus . . . 95
Stimuli and design . . . 95
Procedure . . . 96
Results . . . 97
Within-condition accuracy . . . 98
Within-condition latency . . . 100
Between-condition accuracy . . . 101
Between-condition latency . . . 103
Discussion . . . 103
Absence of between-condition differences . . . 103
Absence of effects in the item analyses . . . 106
No control of meaning dominance . . . 110
5.2 Summary of Problems and Potential Solutions . . . 112
Controlling for relative meaning frequency . . . 112
Selecting a larger, better-controlled item set . . . 113
Employing between- instead of within-participant manipulations . 113
Generalizing across participants and items . . . 114
Powerful difficulty manipulations . . . 114
5.3 Lexical Decision Experiment 2 . . . 114
Methods . . . 116
Participants . . . 116
Apparatus . . . 117
Stimuli and design . . . 117
Procedure . . . 120
Results . . . 121
Accuracy . . . 123

Latency . . . 124
Discussion . . . 126
5.4 General Discussion . . . 129

6 Summary and Discussion of Part I 133

II Towards a General Theory of Response Selection 137

7 Introduction and Brief Review of Computational Models of Response Selection 139
7.1 Theoretical and Methodological Approaches to Studying Response Selection . . 140
The diffusion model . . . 140
Standard connectionist models . . . 144
Connectionist models of response selection processes . . . 146
7.2 Discussion of the Different Approaches and Motivation for the Present Work . . 148

8 Re-evaluating the Empirical Evidence Used to Assess Models of Response Selection: A Behavioral Investigation of Numerosity Judgment 151
8.1 The Ratcliff, Van Zandt, and McKoon (1999) Numerosity Judgement Experiment and Associated Simulations . . . 152
Potential limitations of the study . . . 154
8.2 Numerosity Judgment Experiment . . . 156
Methods . . . 157
Participants . . . 157
Apparatus . . . 157
Stimuli and design . . . 158
Procedure . . . 160
Results . . . 162
Data screening . . . 162
Initial proficiency . . . 163
Overall performance in the constant threshold conditions . . . 164
Sequential effects . . . 172
Accuracy . . . 180
Latency for correct responses . . . 183
Adaptive effects . . . 185
Discussion . . . 191

9 Evaluating an Improved Model of Response Selection: Connectionist Simulations 197
9.1 Developing an Improved Connectionist Formalism . . . 197
9.2 Numerosity Judgement Simulation . . . 200
Methods . . . 201
Simulated participants . . . 201
Network architecture . . . 201
Representations . . . 204

Pre-training and task simulation 205
Results 206
Overall performance in the constant threshold conditions 206
Sequential effects 213
Accuracy 219
Latency for correct responses 220
Adaptive effects 222
Activation trajectories 227
Discussion 229

10 Summary and Discussion of Part II 233

III Summary and Conclusion 239

11 Summary and Conclusion 241

IV Appendixes 245

A SOS! An Algorithm and Software for the Stochastic Optimization of Stimuli 247
A.1 A manual Heuristic for Identifying Optimal Stimuli 249
Problems with the manual heuristic 250
A.2 Stochastic Optimization of Stimuli: A Brief Overview 252
Stochastic optimization 257
A.3 SOS Performance on a Range of Optimization Problems 259
Comparison of SOS stimuli with manually-selected stimuli in a two-level, one-way design 260
A regression analogue and expansion of the one-way design used by Morrison and Ellis (1995) 264
Performance of SOS on additional examples 267
Categorical/ANOVA designs 269
One-way designs 269
Two-way designs 270
Other optimization problems 271
A.4 Extended Applications 271
A.5 Using and Modifying SOS 273
A.6 Conclusion 273
A.7 Subappendix 1: Implemented Cost Functions 274
Soft constraints 274
Simple soft constraints 274
Two-sample distance constraints 274
Single-sample distance constraints 278
Soft entropy constraints 278

Soft correlation-matching constraints 282
Soft meta-constraints 283
Cost-matching meta-constraints 284
Conditional cost-matching meta-constraints 287
Hard constraints 288
Hard bound constraints 289
A.8 Subappendix 2: Tutorial and Details for the ME95 SOS Optimization 291
Creating the samples and population 292
Creating the optimization 294
Adding the constraints 295
Statistical tests 298
Configuring and running the optimization 299
Greedy optimization 300
Stochastic optimization 304
Assessing generalizability 310
A.9 Subappendix 3: Detailed Overview of the Steps in an SOS Optimization 312
Step 1: Defining the Experimental Conditions 312
Step 2: Specifying constraints and penalties for violating the constraints . 312
Step 3: Identifying a population and an initial sample of stimuli 314
Step 4: Search for items in the population and in the sample that could be swapped to better satisfy the constraints 315
Representing the optimization problem as movements in state space315
Greedy optimization of stimuli 320
Greedy optimization's strength is also its weakness—or, the pitfalls of local minima 322
Stochastic optimization of stimuli 324
Adding a temperature annealing schedule for fast, reliable optimizations 328
Using a proxy of thermal equilibrium to assess when temperature should be lowered 332
Step 5. Evaluating how well the constraints have been satisfied and stopping the optimization 334
Step 6. Assessing the degree to which the items in a sample represent the broader population 337

B eDom: Software and Norms for 544 English Homonyms 341
B.1 Issues with Existing Norming Methods 343
B.2 Dictionary Definition Rating Experiment 345
Methods 346
Participants 346
Stimuli 346
Procedure 347
Results 349
Removing data for unknown words 349

Participant supplementation of dictionary definitions 349
Comparison of the meanings listed in the dictionary to those that were familiar to the participants 350
Characteristics of the norms and implications for studies that employ homonymous stimuli 352
Reliability 352
Distribution of largest relative meaning frequency 353
Correlation with other variables 355
Comparison to relative frequency estimates derived from the classification of free associates 358
Predictive validity 359
Analysis of the data reported by Armstrong and Plaut (2011) 359
Using β and U to predict lexical decision performance in the English Lexicon Project 362
Discussion 364
B.3 Conclusion 365

References **367**

List of Figures

2.1 Scatterplot of the number of unrelated meanings and number of related senses associated with 45,005 words in the Wordsmyth Online Dictionary (Parks, Ray, & Bland, 1998). An additional 2,904 words with more than 15 related senses are not depicted. For consistency with later sections of the manuscript, the following labeling scheme was also adopted: words with a single meaning and a single sense were labeled as 'unambiguous', words with only one related sense associated with each unrelated meaning were labeled as 'homonymous', words with multiple related 'senses' associated with a single meaning were labeled as 'polysemous', and words with multiple unrelated meanings, one or more of which was associated with multiple related senses, were labeled as 'hybrid'. 13

3.1 Semantic activation trajectories predicted by the settling dynamics account. The plot shows a measure of semantic activity, f(semantic activation), as a function of settling time in semantics for the contextually appropriate and inappropriate interpretations of polysemous, unambiguous, and homonymous words. The blend-state label serves to delineate the early "context-free" settling dynamics from the later "context-sensitive" settling dynamics. The time point beginning at the y-ordinate corresponds to when semantic processing begins and does not reflect assumed early perceptual and orthographic/phonological processing. . . . 55

4.1 Architecture of the connectionist network from the first settling dynamics simulation. 68

4.2 The average number of semantic units active above 0.7 for polysemous, unambiguous, and homonymous words in the first settling dynamics simulation. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only. No semantic units were active above 0.7 before unit update 10. Slice A: polysemous words settle more quickly than unambiguous words, which in turn settle fractionally more quickly than homonymous words. Slice B: Theoretical cross-over point at which the trajectories for polysemous words and homonymous words are both significantly different form unambiguous words. Slice C: Polysemous words are fractionally faster than unambiguous words, and both are substantially faster than homonymous words. Vertical as opposed to horizontal slices are used because the claim is only that semantics contributes to driving performance in these tasks, not that the current model is a comprehensive account of the dynamics underlying the different behavioral tasks as a whole. 70

4.3 Architecture for the network from the second settling dynamics simulation. Red arrows indicate positive (excitatory) connections and blue arrows indicate negative (inhibitory) connections. The thickness of the arrows indicates the constraints on the magnitudes of the weights that were imposed via weight decay; larger arrows indicate that the weights were less constrained and could be adjusted to larger values. 76

4.4 The number of semantic units with activations above 0.75 for polysemous, unambiguous, and homonymous words in the second settling dynamics simulation. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only. No semantic units were active above 0.75 before unit update 30. Slice A: polysemous words are settling more quickly than unambiguous words, which in turn are settling fractionally more quickly than homonymous words. Slice B: Theoretical cross-over point at which the trajectories for polysemous words and homonymous words are both significantly different from unambiguous words. Slice C: A reversal of the ambiguity advantage occurs; polysemous words are fractionally faster than unambiguous words, and both are faster than homonymous words. Slice D: A processing advantage for unambiguous words is observed relative to both polysemous and homonymous words. 79

4.5 The average amount of semantic activation exceeding a threshold of 0.5 for polysemous, unambiguous, and homonymous words in the second settling dynamics simulation. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only. No semantic units were active above the 0.5 threshold before unit update 19. Slice A: polysemous words are settling more quickly than unambiguous words, which in turn are settling fractionally more quickly than homonymous words. Slice B: Theoretical cross-over point at which the trajectories for polysemous words and homonymous words are both significantly different from unambiguous words. Slice C: Polysemous words are fractionally faster than unambiguous words, and both are faster than homonymous words. Slice D: A processing advantage for unambiguous words is observed relative to both polysemous and homonymous words. 80

4.6 Plot of the incoming weights for one interpretation of each type of word that the second simulation was trained on. The 'x' marks a target unit to which all of the other depicted units send incoming connections. Two plots are presented for the polysemous words: one coresponding to the incoming weights for a core feature shared across senses, and one corresponding to the incoming weights for a distinguishing feature associated with only one of the two senses. The color of the square indicates the sign of the weight: Black squares correspond to inhibition and white squares correspond to excitation. The size of the square corresponds to the magnitude of the weight, with larger squares indicating larger magnitudes. 82

4.7 Summary illustration of the weight diagram presented in Figure 4.6. The distinguishing features of the polyseme appear as a single group of three units and the two distinguishing features are offset. Red indicates positive weights to the group, blue indicates negative weights to the group, and the thickness of the line reflects the approximate magnitude of these weights relative to other between- or within-region weights; the only exception is the context projection to the distinguishing features of the polyseme, which represents only the inputs to the distinguishing features (the incoming weight to the core features was zero). . . . 83

5.1 Lexical Decision Experiment 1: Accuracies for unambiguous, polysemous, and homonymous words for each difficulty condition. Error bars in this and all subsequent figures depict the standard errors of the means. 99

5.2 Lexical Decision Experiment 1: Latencies for unambiguous, polysemous, and homonymous words for each difficulty condition. Significant differences are indicated with asterisks. 102

5.3 Accuracy from the lexical decision experiment that implemented a within-participant difficulty manipulation. Highly significant differences ($p < .005$) are indicated with two asterisks, significant differences ($p < .05$) are indicated with one asterisk, and marginal differences ($p < .15$) are indicated with crosses. 107

5.4 Latency for correct responses from the lexical decision experiment that implemented a within-participant difficulty manipulation. Highly significant differences ($p < .005$) are indicated with two asterisks, significant differences ($p < .05$) are indicated with one asterisk, and marginal differences ($p < .15$) are indicated with crosses. 108
5.5 Plot of the factor loadings for the first component of a factor analysis, labeled fLex, on the SUBTL word frequency and the familiarity data for the different word classes. 111
5.6 Accuracy data for each word class in each condition. E = easy nonwords. E-F = easy-full. H-F = hard-full. E-D = easy-degraded. H-D = hard-degraded. P-F = pseudohomophone-full. H = homonym. U = unambiguous. P = polyseme. Y = hybrid. NW = nonword. Significant ($p < .05$) and marginal ($p < .1$) differences between homonyms, polysemes, and hybrid items relative to unambiguous items are denoted by single and double lines, respectively. Note that these differences were calculated in a multiple regression that included a number of additional covariates. Consequently, the tests are not directly comparing the raw condition means presented here. 125
5.7 Latency data for each word class in each condition. E = easy nonwords. E-F = easy-full. H-F = hard-full. E-D = easy-degraded. H-D = hard-degraded. P-F = pseudohomophone-full. H = homonym. U = unambiguous. P = polyseme. Y = hybrid. NW = nonword. Significant ($p < .05$) and marginal ($p < .1$) differences between homonyms, polysemes, and hybrid items relative to unambiguous items are denoted by single and double lines, respectively. Note that these differences were calculated in a multiple regression that included a number of additional covariates. Consequently, the tests are not directly comparing the raw condition means presented here. 127

8.1 Plot of accuracy as a function of the distance in number of asterisks from the response threshold. 167
8.2 Plot of latency as a function of the distance in number of asterisks from the response threshold. Separate lines show the data for correct and incorrect responses. 168
8.3 Plot of latency as a function of accuracy generated by combining the results from Figures 8.1 and 8.2. Separate lines show the data for correct and incorrect responses. 169
8.4 Distribution of latencies for correct and incorrect responses. 170
8.5 Degree of skew in the distribution of latencies for correct responses as a function of distance in number of asterisks from the threshold. 171
8.6 Accuracy as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks. 174
8.7 Latency for correct responses as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks. 175
8.8 Boxplots of the latency data for correct responses for the first 400 trials in each condition. 176

8.9 Accuracy as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines. . 177
8.10 Latency for correct responses as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines. 178
8.11 Accuracy as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions. 187
8.12 Latency for correct responses as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions. 188
8.13 Accuracy as a function of trial rank. Separate panels present different groupings of standardized number of asterisks and separate lines within a plot represent the different conditions. 189
8.14 Latency for correct trials as a function of trial rank. Separate panels present different groupings of standardized number of asterisks and separate lines within a plot represent the different conditions. 190

9.1 Architecture for the network from the numerosity judgement simulation. Red arrows indicate positive (excitatory) connections and blue arrows indicate negative (inhibitory) connections. 203
9.2 Plot of accuracy as a function of the distance in number of asterisks from the response threshold. 208
9.3 Plot of latency as a function of the distance in number of asterisks from the response threshold. Separate lines show the data for correct and incorrect responses. 209
9.4 Plot of latency as a function of accuracy generated by combining the results from Figures 9.2 and 9.3. Separate lines show the data for correct and incorrect responses. 210
9.5 Distribution of latencies for correct and incorrect responses. 211
9.6 Degree of skew in the distribution of latencies for correct responses as a function of distance in number of asterisks from the threshold. 212
9.7 Accuracy as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks. 215
9.8 Latency for correct responses as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks. 216
9.9 Accuracy as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines. . 217
9.10 Latency for correct responses as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines. 218
9.11 Accuracy as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions. 224

9.12 Latency for correct responses as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions. 225

9.13 Latency for correct trials as a function of trial rank. Separate panels present different groupings of standardized number of asterisks and separate lines within a plot represent the different conditions. 226

9.14 Activation trajectories for the 'low' (black) and 'high' (red) response units. Both plots show the correct activation of the 'low' response unit following the presentation of a 'low' number of asterisks. The top panel shows the trajectories associated with the presentation of a stimulus that was far from the response threshold (32 asterisks). The bottom panel shows the trajectories associated with the presentation of a stimulus that was close to the response threshold (44 asterisks). Note that different y-axis scales were used for each plot to highlight the relative similarity between both plots. 228

A.1 Depiction of the likelihood of "moving" to the "swap" item or "staying" with the current item based on a range of values of the cost difference, $\Delta cost$, between the two items that could be swapped, and temperature. Inf = infinite. 258

A.2 Plots generated by the SOS software of the pre- and post-optimization AoA and frequency distributions when a constraint was imposed to pressure the creation of uniform distributions across the populations (AoA range: 125-694; frequency range: 1-214). The pre-optimization AoA and frequency distributions had entropy values of -0.012 and -0.071, and these entropy values increased to -0.0095 and -0.0061 by the end of the optimization. These final distributions did not differ from a uniform distribution ($p > .001$). 268

A.3 The main window of the SOS graphical user interface. The process of running an optimization generally flows from top to bottom, left to right. 292

A.4 Plot generated by the SOS software of values for several metrics throughout the course of the greedy optimization. The topmost plot displays temperature, the second plot displays cost, the third plot displays $\Delta cost$, the fourth plot displays the likelihood of swapping to a neighboring item, p(swap), and the last plot displays the p-values for the statistical tests. 305

A.5 Plot generated by the SOS software of values for several metrics throughout the course of the stochastic optimization. A more detailed description of the plots is included in the caption for Figure A.4. 311

A.6 Upper section: Two-dimensional depiction of the state space formed by each of the different combinations of pairs of items listed in Table A.2. The values in each cell of this table correspond to the cost associated with each combination of items when trying to minimize the variance in a sample of two items. The diagonal is blank due to the restriction that items are sampled without replacement. Lower section: A reduced version of the table from the upper section where the data from the upper triangle has been shifted to the left to merge with the lower triangle. The labels for the vertical axis of this table remain unchanged, whereas the labels for the horizontal dimension now correspond to the neighboring items of the items listed on the vertical axis. 318
A.7 Three-dimensional representation of the data from the lower section of Figure A.6. The plane at the base of the plot corresponds to the different combinations of items that can be formed from the data listed in Table A.2. The "item" axis corresponds to items 1-4 from the table directly, whereas the "neighbor" axis corresponds to the neighbor items of the item selected on the "item" axis. The "cost" axis represents the cost associated with each state, or combination of items, that could be formed from the population of items being sampled from. Note that the shape of the cost surface is concave when viewed from above and consists of a minimum valley surrounded by hills of monotonically increasing cost. The bottom of this valley corresponds to the global minimum of the state space, and to the pair of items with the lowest variability in frequency values. 319
A.8 Depiction of a state space containing both a local minimum (foremost region of the plane) and a global minimum (backmost region of the plane) separated from one another by a 'hill' of higher cost. The cost function used to generate the space was identical to that used for Figure A.7, applied to the data in Table A.3. . 323
A.9 Approximation of the state space illustrated in Figure A.8 after a multidimensional scaling has been applied to the base plane on that figure to reduce it to a single dimension. 324
A.10 Depiction of the likelihood of moving to a neighboring sample in state space by swapping two items based on the cost difference, $\Delta cost$, between the two items that could be swapped, for a range of temperature values. Inf = infinite. 326
A.11 Depiction of the temperature values after a given number of temperature steps away from an initial temperature, T_0, as a function of a range of pDecrease and λ values. 331

B.1 Screen capture of the *eDom* norming software during the norming of the word PUPIL. 350
B.2 Standard error of the largest meaning frequency estimates for each normed word as a function of each word's largest meaning frequency. The solid line depicts the mean of the standard error for all words with the same largest meaning frequency, fit with a cubic smoothing spline. 354
B.3 Histogram of the distribution of the largest relative meaning frequency value for all of the normed words. 356

List of Tables

5.1 Lexical Decision Experiment 1: Word and Nonword Positional Bigram Frequencies 97
5.2 Lexical Decision Experiment 1: Accuracy 99
5.3 Lexical Decision Experiment 1: F-statistics for the 2 × 2 Within-subjects ANOVAs 101
5.4 Lexical Decision Experiment 1: Latency 102
5.5 Lexical Decision Experiment 2: Properties of Word Stimuli 119
5.6 Lexical Decision Experiment 2: Properties of Nonword and Word Stimuli 120
5.7 Lexical Decision Experiment 2: Accuracy 126
5.8 Lexical Decision Experiment 2: Latency 128

8.1 Statistics for the predictors in the accuracy analysis 182
8.2 Statistics for the predictors in the analysis of the latencies for correct trials 184

9.1 Statistics for the predictors in the accuracy analysis 221
9.2 Statistics for the predictors in the analysis of the latencies for correct trials 221

A.1 Descriptive statistics for the samples used in Experiment 2 of Morrison and Ellis (1995) and for the SOS optimized samples 290
A.2 Example ‘population’ data set containing a single global minimum when attempting to minimize the variability in the frequencies of a ‘sample’ containing two items 316
A.3 Example ‘population’ data set containing a global minimum and a local minimum when attempting to minimize the variability in the frequencies of a ‘sample’ containing two items 323

Chapter 1

Introduction and Overview

The vast majority of words in English and in other languages are semantically ambiguous and their meanings vary depending on the context in which they occur (e.g., the <river> vs. the <money> meanings of the word BANK). Consequently, understanding ambiguous word processing is a critical component of any comprehensive theory of word and discourse comprehension. Nevertheless, a comprehensive account of semantic ambiguity has not yet been produced. Arguably, this is primarly due to two factors: the apparent inconsistencies in the effects reported both within and across tasks (e.g., Frazier & Rayner, 1990; Hino, Pexman, & Lupker, 2006; Rodd, Gaskell, & Marslen-Wilson, 2002), and the tendency to propose narrow (and often task-specific), underspecified, and/or inconsistent accounts of extant ambiguity effects that do not help reconcile the task differences or guide future research (e.g., Van Petten & Kutas, 1987; Hino, Kusunose, & Lupker, 2010; Hino & Lupker, 1996; Hino et al., 2006; Rodd, Gaskell, & Marslen-Wilson, 2004).

The work reported in this dissertation served primarily to develop a broader and more comprehensive theory of semantic ambiguity resolution based on the domain-general learning, representation, and processing assumptions of the connectionist framework (McClelland & Rumelhart, 1981; Rumelhart & McClelland, 1986a, 1986b; Rumelhart, Durbin, Golden, & Chauvin, 1995). This theory is refered to as the 'settling dynamics' account and is the primary focus of

Part I. In conducting this work, it also became apparent that the simplifying assumptions that typically segregate the lexical processing system from the system that generates responses in many tasks used to study semantic ambiguity precluded the evaluation of certain aspects of the settling dynamics account and other competing theories. Part II evaluates the capacities and assumptions of several response selection models (e.g., Joordens, Piercey, & Azarbehi, 2009; Ratcliff, 1978; Usher & McClelland, 2001) and, finding these models to be lacking in several respects, proposes a new approach to studying response selection based on the same theoretical framework used to explain the ambiguity effects. The proposed theories are supported by connectionist modeling that serves to explicitly illustrate and explore the implications of the theories, and by behavioral investigations that evaluate the assumptions and predictions of the accounts. Part III consists of a general discussion of how an amalgamation of these two bodies of work can lead to a deeper understanding of semantic ambiguity, response selection, and cognitive processing more generally. Additional methodological advances developed as part of this work that relate to stimulus selection and the norming of relative frequencies of an ambiguous word's meanings are reported in the Appendixes presented in Part IV. An overview of the chapters in each section follows.

Chapter Overview

Part I. Towards a General Theory of Semantic Ambiguity Resolution

Chapter 2. Introduction and Review of the Semantic Ambiguity Literature. The second chapter introduces the semantic ambiguity literature and reviews the results of numerous studies that have employed a wide range of tasks and methodologies. Subsequently, several accounts of semantic ambiguity effects are reviewed and, being found lacking, the characteristics of an improved account are identified to guide the subsequent work.

Chapter 3. The Settling Dynamics Account. On the basis of the insights gleaned from the literature review, the third chapter introduces a theoretical account of ambiguous word comprehension that captures a wide range of semantic ambiguity effects. The overarching premises of this theory are that the range of semantic ambiguity effects that have been reported are due to two main factors: 1) the temporal dynamics underlying a) co-operation and competition amongst an ambiguous word's semantic representations and b) contextual integration, and 2) the overall precision of the semantic code needed to make a decision in a given task. Together, these dynamics cause various ambiguity effects to emerge at different points in time within a system instantiating a single neurobiologically-inspired computational formalism. The first section of the chapter introduces the account in more detail, the second section describes how the account captures, in principle, a range of ambiguity effects, and the third section describes key challenges to this account and how they can be addressed via computational and behavioral investigations.

Chapter 4. Substantiating the Settling Dynamics Account: Connectionist Simulations. The fourth chapter reports two connectionist simulations aimed at demonstrating how the fundamental assumptions of the settling dynamics account can produce a portion of the theorized semantic settling dynamics outlined in the previous chapter. In particular, this work focuses on the more contentious early portion of the settling dynamics that occur in the absence of strong biasing context (Hino et al., 2006). The first simulation was conducted using a standard set of connectionist assumptions and produced a reasonable—although not perfect—approximation of the theorized semantic settling dynamics. The second simulation employed a more neurobiologically plausible set of connectivity assumptions and the simplest possible representation assumptions that were assumed to be needed to produce the expected settling dynamics. This simulation produced an improved set of semantic settling dynamics and allowed for cleaner insight into how the underlying mechanics of the simulation led to the observed dynamics. Following the presentation of these simulations, the general strengths and weaknesses of this work are discussed and directions for future simulation work are identified.

Chapter 5. Evaluating the Settling Dynamics Account: Behavioral Investigations. The work described in the fifth chapter evaluates the settling dynamics account by testing one of its key predictions: Variations in the degree of semantic precision required to generate a response in a given task will determine the magnitudes of the different ambiguity effects that are observed. This work is conducted in the context of lexical decision tasks that are extensions of a design previously employed by Rodd et al. (2002). The results of the first set of experiments were, at first glance, in strong agreement with the predictions of the settling dynamics account. However, more detailed analyses of the data identified several issues with this study that forced this conclusion to be rescinded. To address these issues, a stimulus optimization algorithm and software, along with new methods to norm the relative frequency of ambiguous words' meanings, were developed and used to design an improved item set. A lexical decision task that employed this item set and attemped to vary processing time via both nonword manipulations (orthographically wordlike nonwords vs. very orthographically wordlike nonwords vs. pseudohomophones) and stimulus contrast manipulations (stimuli presented as white-on-black vs. grey-on-black) produced effects that were broadly (albeit not perfectly) consistent with the settling dynamics account.

Chapter 6. Summary and Discussion of Part I. The sixth chapter summarizes and discusses the successes and limitations of the work that was done to assess the validity of the settling dynamics account. In doing so, one important direction for additional work is identified—understanding how the semantic system could be integrated with a response selection system to model specific tasks, directly compare competing theories, and permit the emergence of new behavior that could explain additional phenomena that have been reported in the literature.

Part II. Towards a General Theory of Response Selection

Chapter 7. Introduction and Brief Review of Computational Models of Response Selection. The next major advance of the settling dynamics account would involve the integration of a model of orthographic, phonological, and semantic processing with a model of response selection. This will permit the modeling of actual tasks and the direct testing and comparison of different theories of semantic ambiguity effects. The seventh chapter examines how the response selection literature can inform the development of a model of response selection that is suitable for integration with the work presented in Part I. The first section reviews the diffusion model and two classes of models based on variants of the connectionist framework that have been employed to understand response selection, so as to identify their strengths and weaknesses. In so doing, this work identifies the key computational principles and supporting empirical data that are needed to develop an improved domain-general connectionist model of response selection which can, amongst many other potential uses, be integrated with the settling dynamics account.

Chapter 8. Re-evaluating the Empirical Evidence Used to Assess Models of Response Selection: A Behavioral Investigation of Numerosity Judgment. The eighth chapter aims to assess the validity of the behavioral results reported by Ratcliff et al. (1999) as a 'gold standard' for model comparison in general, and as a strong demonstration that the connectionist framework is not suitable for modeling response selection in particular. This is accomplished first by considering whether certain esoteric aspects of the Ratcliff et al. study may have led to atypical results that are not in line with the fundamental assumptions of connectionist models (e.g., commitment to error-driven learning). An extension of their original study is then run to evaluate whether minor changes in the characteristics of the task that make it more representative of standard experiments leads to major alterations in participants' performance that are more consistent with the predictions of connectionist accounts. The implications of this work with respect to the evaluation and development of connectionist (and potentially other) models of response selection

are also briefly discussed.

Chapter 9. Evaluating an Improved Model of Response Selection: Connectionist Simulations. Having established a new set of data using a task that is more representative of standard psychological experiments, the ninth chapter aims to re-evaluate how well a model of response selection based on a more biologically plausible connectionist formalism fares at capturing the behavioral effects. The results show that the model does suprisingly well and captures many of the overall, sequential, and adaptive effects from the behavioral data that are outside the scope of standard models of response selection. Taken together, this work supports the use of a biologically plausible connectionist framework to simulate response selection, both in two-alternative forced choice tasks and beyond. Possible improvements on this work that are motivated by some minor detailed failures of the simulations are also discussed.

Chapter 10. Summary and Discussion of Part II. The tenth chapter summarizes and discusses the success of the numerosity judgment experiment and associated connectionist simulations at undermining the validity of the claims of Ratcliff et al. (1999) that the connectionist framework is ill-suited for studying response selection. Additional implications for future investigations of response selection, including the extension of these investigations to a broader range of more complex tasks and types of possible responses, are also discussed in relation to current practices in the field.

Part III. Summary and Conclusion

Chapter 11. Summary and Conclusion. The final chapter summarizes the dissertation work and draws some general conclusions about its implications for future research.

Appendix A. SOS! An Algorithm and Software for the Stochastic Optimization of Stimuli. Appendix A presents SOS, an algorithm and software package for the Stochastic Optimization of

Stimuli that was developed in collaboration with Christine Watson and David Plaut (Armstrong, Watson, & Plaut, 2012). SOS takes its inspiration from a simple manual stimulus selection heuristic that has been formalized and refined as a stochastic relaxation search. The algorithm rapidly and reliably selects a subset of possible stimuli that optimally satisfy the constraints imposed by an experimenter. This allows the experimenter to focus on selecting an optimization problem that suits his or her theoretical question and to avoid the tedious task of manually selecting stimuli. In doing so, the algorithm facilitates research using factorial, multiple/mixed-effects regression, and other experimental designs. The use of SOS is demonstrated with a case study and other research situations that could benefit from this tool are discussed. Support for the generality of the algorithm is demonstrated through Monte Carlo simulations on a range of optimization problems faced by psychologists. The software implementation of SOS and a user manual are provided free of charge for academic purposes as pre-compiled binaries and MATLAB source files at `http://sos.cnbc.cmu.edu`.

Appendix B. eDom: Software and Norms for 544 English Homonyms. When using homonyms as stimuli, it is critical to control for the relative frequencies of each interpretation, because this variable can drastically alter the empirical effects of homonymy. Currently, the standard method for estimating these frequencies is based on the classification of free associates generated for a homonym, but this approach is both assumption-laden and resource-demanding. Appendix B presents an alternative norming methodology based on explicit ratings of the relative meaning frequencies of dictionary definitions that was developed with Natasha Tokowicz and David Plaut (Armstrong, Tokowicz, & Plaut, 2012). To evaluate this method, normative data were collected and analyzed for 544 English homonyms using the eDom norming software that was developed for this purpose. The results showed that dictionary definitions were generally sufficient to exhaustively cover word meanings, and the methods converged on stable norms with fewer data and less effort on the part of the experimenter. The predictive validity of the norms was demonstrated in analyses of lexical decision data from the English Lexicon Project

(Balota et al., 2007), and from Armstrong and Plaut (2011). On the basis of these results, this new norming method obviates relying on the unsubstantiated assumptions involved in estimating relative meaning frequencies on the basis of the classification of free associates. Additional details related to the norming procedure, the meaning frequency norms, and the source code, standalone binaries, and user manual for the software are available at `http://edom.cnbc.cmu.edu`.

Part I

Towards a General Theory of Semantic Ambiguity Resolution

Chapter 2

Introduction and Review of the Semantic Ambiguity Literature

Developing a theory of how words are represented and processed is important both for understanding the psychological underpinnings of language, and for gaining insight into the representations and processing mechanisms underlying cognition more generally (McClelland & Rumelhart, 1981; Plaut, McClelland, Seidenberg, & Patterson, 1996). Much influential work on this issue has focused on problems such as the relationship between orthography and phonology and has largely ignored or made gross simplifications regarding the contributions from semantics (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2000; Plaut et al., 1996). This work, although clearly of great value for other reasons, nevertheless does little to explain one of the most fundamental objectives of language—how orthographic and phonological 'surface' forms of words are mapped onto 'deeper' semantic representations of a word's meaning.

Developing a theory of how words are comprehended[1]—that is, of how surface and deep

[1]The literature sometimes refers to this process as one of word recognition (e.g., Plaut, 1997), but 'comprehension' is used in the present work to distinguish between the process of recognizing a suface representation of a whole word (i.e., a lexical entry, Coltheart et al., 2000), if such a representation exists, from that of recognizing the meaning associated with a surface form.

representations are mapped to one another— is not, however, a simple undertaking. One of the key reasons for this is that a single word is usually associated with multiple different interpretations depending on the context in which it occurs. For instance, BANK may refer to a financial institution or to the edge of a body of water, and PAPER may refer to an academic work or the material loaded into a printer (for brevity of notation, these semantic ambiguities are denoted later in the form of <river>/<money> BANK). In fact, ambiguous words such as these make up the vast majority of the content words in English and other languages (Klein & Murphy, 2001). This is illustrated in the scatterplot presented in Figure 2.1, which presents the number of separate interpretations associated with the words in the Wordsmyth Online Dictionary (Parks et al., 1998). This scatterplot is broken down on the basis of whether those interpretations were 'senses' that were related to one another (e.g., the interpretations of PAPER, which may all share some common features regarding the physical characteristics of the object to which they refer) or 'meanings' that were unrelated to one another and did not share a core interpretation (e.g., the interpretations of BANK), according to the Wordsmyth lexographers. These measures have also been shown to correlate with human mental representations of these words (Azuma & Van Orden, 1997; Rodd et al., 2002). The scatterplot clearly shows that a large number of words are not unambiguous and associated with only a single interpretation; rather, many words are associated with multiple interpretations. By one estimate, these ambiguous words represent over 80% of frequently-used content words in English (Klein & Murphy, 2001). Nevertheless, these ambiguous words typically go unnoticed during comprehension because of the proficiency of the comprehension system at coping and resolving these ambiguities. Developing a theory of semantic ambiguity (sometimes called 'lexical' ambiguity) is therefore critical both to understanding word comprehension in isolation, and in moving towards a more general account of contextually-influenced comprehension of multiple words (e.g., McClelland, St John, & Taraban, 1989).

In response to the theoretical importance of understanding semantic ambiguity resolution,

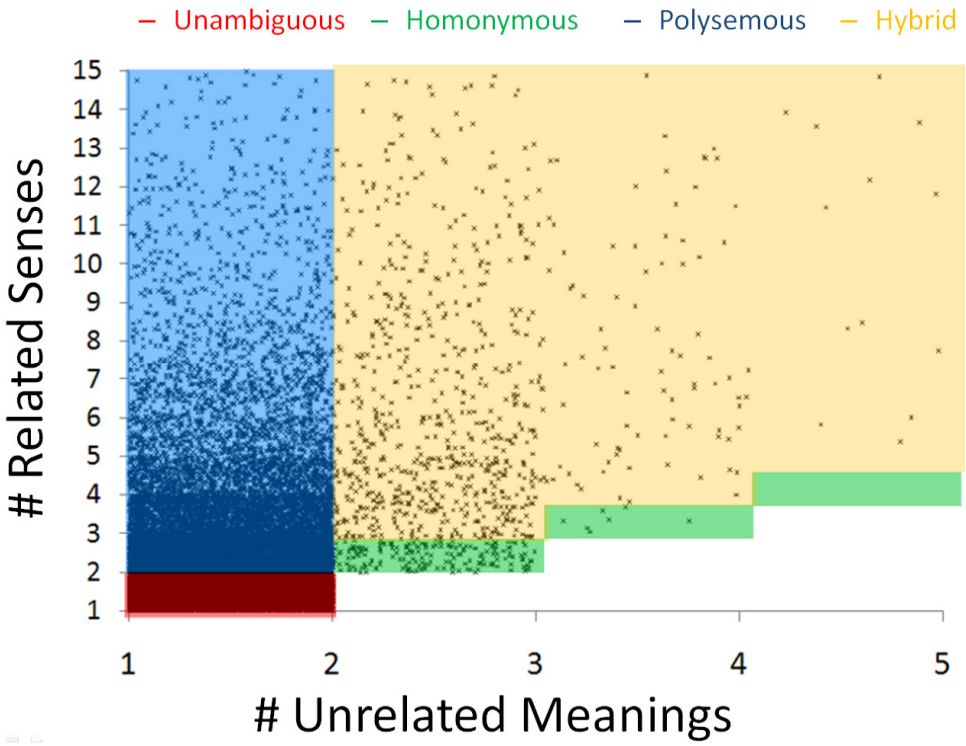


Figure 2.1: Scatterplot of the number of unrelated meanings and number of related senses associated with 45,005 words in the Wordsmyth Online Dictionary (Parks et al., 1998). An additional 2,904 words with more than 15 related senses are not depicted. For consistency with later sections of the manuscript, the following labeling scheme was also adopted: words with a single meaning and a single sense were labeled as 'unambiguous', words with only one related sense associated with each unrelated meaning were labeled as 'homonymous', words with multiple related 'senses' associated with a single meaning were labeled as 'polysemous', and words with multiple unrelated meanings, one or more of which was associated with multiple related senses, were labeled as 'hybrid'.

a large body of research has been devoted to its study over the past 40 years (e.g., Azuma & Van Orden, 1997; Beretta, Fiorentino, & Poeppel, 2005; Borowsky & Masson, 1996; Frazier & Rayner, 1990; Gernsbacher, 1984; Hino et al., 2010; Hino & Lupker, 1996; Hino et al., 2006; Jastrzembski, 1981; Joordens & Besner, 1994; Klein & Murphy, 2001; Klepousniotou & Baum, 2007; Millis & Button, 1989; Mirman, Strauss, Dixon, & Magnuson, 2010; Pexman & Lupker, 1999; Rodd et al., 2002; Rubenstein, Garfield, & Millikan, 1970; Schvaneveldt, Meyer, & Becker, 1976). This work has typically reported substantial performance differences between semantically ambiguous words and unambiguous controls. However, the impact of this work has been quite limited because of the considerable inconstencies in the results reported both within and between tasks, and the often unintuitive effects of number and relatedness of an ambiguous word's interpretations—not to mention the methodological issues that undermine many studies. Developing a theory that subsumes these varied effects and that identifies the mechanisms underlying these discrepancies is therefore a necessary step in leveraging these studies towards theories of word and discourse processing, and beyond.

The following sections review key data and theories that have emerged from the semantic ambiguity literature that are relevant to the development of a more comprehensive account of the mechanisms underlying ambiguous word comprehension. It is worth noting, however, that the order of presentation of this review and of the later description of the proposed account of these effects does not reflect the order in which all of this work was completed; rather, both of these bodies of work completed in parallel. Consequently, the theory outlined in the next chapter should not be viewed as an entirely post hoc explanation of these data.

2.1 Review of Key Results from the Semantic Ambiguity Literature

The subsequent sections review key results from the semantic ambiguity literature. The goal of this review is to identify the typical patterns of results that have been observed across a range of tasks. As part of this process, likely causes for some of the discrepant results that have been reported—particularly, but not exclusively, across highly similar tasks—are discussed. The studies have been organized in terms of whether they involve or do not involve context that intentionally biases the activation of a particular interpretation of an ambiguous word. This grouping has several advantages. First, it distinguishes between the studies that could require a more complex theoretical account that relates to the selection of a particular interpretation and the studies that may be explained via simpler accounts (e.g., Swinney, 1979). Second, in general this grouping also tends to cluster together tasks that are associated with rapid processing (tasks that do not involve context) and tasks that are associated with slower processing (tasks that do involve context). Third, this grouping reflects the existing organization of much of the literature, in which researchers tend to focus on either the details of rapid ambiguous word processing or on contrasting the different effects observed in the presence or absence of strongly biasing context. Of course, this grouping is not absolute given the nature of several past studies, and those that bridge this gap are presented in the most appropriate section on the basis of other related work.

Before proceeding further, however, several issues should be addressed up-front. To begin, it is worth noting two main factors that often influence ambiguity resolution and how they will be referred to throughout the dissertation. The first is the number of interpretations associated with a word—that is, whether the word is unambiguous or ambiguous. The second is the relatedness amongst the interpretations of an ambiguous word. Although the degree of relatedness is typically assumed to vary continuously (Klepousniotou & Baum, 2007; Klepousniotou, Titone, & Romero, 2008), to remain consistent with the terminology used in past work it is useful to

subdivide this continuous measure of ambiguity into two main groups of items: items with relatively unrelated interpretations, termed *homonyms*, and items with related interpretations, termed *polysemes*; the latter of which account for the vast majority of all ambiguous words (Beretta et al., 2005; Klein & Murphy, 2001). Items may also be both polysemous and homonymous; these are refered to as *hybrid* items.

Another important issue relates to the statistical methods that have been employed in many past studies. In an optimal setting, psycholinguistic experiments should aim to generalize both across participants and across items, and employ appropriate methods and analytical techniques to make such generalizations possible (Baayen, Davidson, & Bates, 2008; Clark, 1973; Raaijmakers, Schrijnemakers, & Gremmen, 1999). When this does not appear to be possible, some researchers have proposed that it is sufficient to provide very close matching of the items across conditions and only seek to generalize the results observed in a given task across participants, leaving generalization across items unaddressed in statistical terms (Hargreaves, Pexman, Pittman, & Goodyear, 2011; Hino & Lupker, 1996; Wike & Church, 1976). This last proposal is not considered to be acceptable in the present work because it is viewed as fundamentally undermining a critical goal of the enterprise–making general claims about psycholinguistic processes—and can lead to misleading interpretations of an experiment's results (see Chapter 5). Researchers often fail to control for or regress out important confounding variables, such as the length (e.g., in number of syllables), familiarity, and the number and relatedness of the interpretations associated with a word. This, in turn, has led to several studies that have called a number of purported ambiguity effects into question (Cutler, 1981; Gernsbacher, 1984; Joordens & Besner, 1994; Mehler, Segui, & Carey, 1978; J. E. Newman & Dell, 1978; Rodd et al., 2002). Indeed, studies that employ appropriate item-level control and statistical methods for generalizing across participants and items are few and far between, and often the data needed to re-analyze these results in a more rigorous manner are unavailable and only an extremely conservative re-analysis is possible (e.g., F-max; Clark, 1973). Consequently, conducting this review required choosing to either re-

evaluate and underpower the results obtained in many of these experiments or accept the results at face value with caveats, as appropriate. Given the small magnitude of many ambiguity effects, the former option was chosen, although certain studies that conflated ambiguous words with related and unrelated interpretations are excluded because of the central role of these conditions in later theory development.

Finally, in most of these studies data for both accuracy and latency for correct responses were reported, but the accuracy data were typically near ceiling. As such, the effects of ambiguity tended to emerge in the latency data. When effects were observed in the accuracy data they tended to recapitulate the effects observed in the latency data. For instance, a processing advantage in the latency data (e.g., faster responses) was also sometimes detected in the accuracy data (e.g., more accurate responses). More critically, the accuracy effects did not run in the opposite direction to the latency effects, as would be expected if the ambiguity effects observed in the latency data were due to some form of speed-accuracy trade-off. Consequently, a full discussion of the accuracy data reported in these studies has been omitted except when it is the primary measure in which a particular effect was reported.

Studies that do not involve context

Many tasks have been used to study semantic ambiguity (for past reviews, see Simpson, 1984, 1994), each of which has its strengths and weaknesses in terms of how natural, how complex, and how pure a measure of semantic processing it is (i.e., the degree to which performance in the task could be influenced by pre- or post-lexical processes in addition to semantic influences; for discussion, see Simpson, 1994). One of the most frequently employed tasks for investigating semantic ambiguity in particular and lexical processing in general is the lexical decision task (e.g., Borowsky & Masson, 1996; Jastrzembski, 1981; Klepousniotou & Baum, 2007; Millis & Button, 1989; Rodd et al., 2002; Rubenstein et al., 1970) and a large portion of this section is devoted to this task. In a very abstract sense, this task may be thought of as assessing some aspect

of how ambiguous words activate semantic representations when there is no systematic pressure to select a particular interpretation of the word, provided that the nonword foils are sufficiently word-like to rule out surface-level discriminations (for alternative views, see Balota & Chumbley, 1984; Hino & Lupker, 1996; Seidenberg, Waters, Sanders, & Langer, 1984). Insofar as a lexical decision can be made based on very coarse semantic information (i.e., is a given character string associated with any meaning whatsoever), as is generally suggested on the basis of the rapid latencies typically associated with responses in this task, the lexical decision task may also be generally thought of as probing relatively early semantic access (Piercey & Joordens, 2000; Masson & Borowsky, 1995).

A large number of isolated-word lexical decision studies of ambiguity effects have been run in the past (i.e., studies that present sequences of single-word trials, with no relationship between the words in the sequence). These studies have typically reported an overall advantage for ambiguous words. However, most of these studies were shown to either be confounded with other factors that influence lexical decision performance (see Gernsbacher, 1984), or collapsed across both homonymy and polysemy and reported only an overall effect of ambiguity which need not be representative of both classes of items (e.g., Borowsky & Masson, 1996; Hino & Lupker, 1996; Kellas, Ferraro, & Simpson, 1988; Millis & Button, 1989; although Hino et al., 2010, argue that the earlier Hino & Lupker study does not suffer from this confound). The present review is therefore primarily restricted to studies that address these issues in their designs. An influential study by Rodd et al. (2002), illustrates the necessity of restricting the review in this manner (see also the earlier related but less-well controlled work by Jastrzembski, 1981, and by Azuma & Van Orden, 1997). The overarching aim of their work was to evaluate the effects of relatedness amongst interpretations in producing the ambiguity advantage that had been reported in previous lexical decision studies; a distinction that had been noted in several prior investigations but that has largely been ignored in the broader literature (e.g., Frazier & Rayner, 1990; Jastrzembski, 1981; Klein & Murphy, 2001). Rodd and colleagues reviewed previous work by Millis

and Button (1989), Borowsky and Masson (1996), and Azuma and Van Orden (1997), in which the authors reported a processing advantage for ambiguous words, to determine whether these effects were caused by polysemes with related senses or homonyms with unrelated meanings. They found that the ambiguous items in these studies consisted primarly of polysemes, and often did not differ from unambiguous items in terms of their number of unrelated meanings. This suggested that the frequently reported ambiguity advantage was actually a more restricted polysemy advantage. A re-interpretation of the Piercey and Joordens (2000) lexical decision experiment that employed the same experimental items as the Borowsky and Masson (1996) experiment further supports this possibility—it also showed an early polysemy advantage (originally reported as an ambiguity advantage) in the absence of contextually-biasing context.

To explore this novel finding, Rodd et al. (2002) conducted three lexical decision experiments; the first two used a visual lexical decision paradigm and the third used an auditory lexical decision paradigm. The first experiment included a large set of polysemous, homonymous, and unambiguous items and regressed out the effects of polysemy (i.e., number of senses associated with a word) and homonymy (i.e., number of meanings associated with a word) in the data analysis. The results indicated that there was a disadvantage for processing homonymous words, such that homonyms were responded to more slowly than either polysemes or unambiguous controls.[2] The second experiment crossed the number of unrelated meanings (one/many[2-3 meanings]) and the number of related senses (few[5 senses]/many[15 senses]) associated with the word stimuli in a factorial design. As it is relevant to a number of later studies, it is worth noting that the total number of related senses in the few and many senses conditions was either loaded onto a single meaning in the case of one-meaning words, or split (potentially unequally) across multiple meanings in the many-meanings conditions. Consequently, the presence of a main effect of meaning was not confounded with the total number of senses associated with a word (cf. Hino et al., 2010). The 'many senses' conditions also had relatively high numbers

[2]Rodd et al. (2002) also report a polysemy advantage in their first experiment, but this is relative to the homonyms and not to the unambiguous words.

of senses (15), which may have further enhanced the polysemy effects, because later work has shown that ambiguity effects increased in magnitude as a function of the number and connectedness amongst interpretations (Locker, Simpson, & Yates, 2003, although a lack of a baseline condition in that work does not allow for the direct assessment of whether an increased number of senses facilitated processing or a reduced number impeded processing).

The results showed a significant processing advantage for words with many senses (i.e., faster latencies) and very small numerical trends towards a processing disadvantage for words with many meanings (i.e., slower latencies). These effects were all largest when pseudohomophones as opposed to word-like nonwords were employed (see also Azuma & Van Orden, 1997); in the latter case, no effects of ambiguity were observed. Thus, although the second experiment agreed with the first in terms of the rank ordering of the conditions, there was no agreement in terms of the significance of the meaning disadvantage or of the sense advantage. This was surprising in many respects given the many similarities in the items and methods used in the two tasks, as well as the highly similar overall performance (responses were only 12 ms faster in the first experiment than in the second experiment). However, the first experiment was exceptional in showing a significant effect of concreteness, albeit using items that spanned a greater range of concreteness values than in the other two experiments reported in that paper (as determined in a re-analysis of these items conducted by Armstrong & Plaut, 2011). This raises the possibility that there was a difference in the overall strength of the semantic influence in the results of the first task in addition to the differences in the ambiguity effects. This may have occured for a number of reasons, including poor control over confounding variables or change variability in the neural activity evoked during semantic access and how it drives the response system (discussed in more detail in the context of the Beretta et al., 2005 study). Additionally, the failure to observe a polysemy advantage in the first experiment may be due to the relatively small difference in the mean number of senses between the polysemous and unambiguous items, which may have underpowered the sense effects (although another study by Klepousniotou & Baum, 2007, has shown that

a polysemy effect can, in principle, be detected with this difference in number of senses between conditions). Consequently, a conservative interpretation of these results and of the Rodd et al. (2002) reanalyses of previous studies would be that visual lexical decision produces a weak and inconsistent homonymy disadvantage, and a somewhat more reliable polysemy advantage.

Using the same type of design and a subset of the word stimuli from the second experiment, a third experiment in the auditory domain produced a similar but substantially stronger pattern of results: both a significant disadvantage for words with many meanings (i.e., slower latencies) and a significant advantage for words with many senses (i.e., faster latencies) were detected. Of relevance to the later behavioral experiments reported in Chapter 5, it is also worth noting that this was accomplished without using pseudohomophones (which do not exist in the auditory domain). Additionally, the overall latencies in the auditory task were approximately 350 ms slower than those in the analogous visual task—a fact that Rodd et al. (2002) attribute at least in part to the length of time needed to present each auditory stimulus. Consequently, although visual and auditory lexical decision are highly similar in many respects, the large latency difference between these tasks suggests that the results of these two tasks should not be grouped without understanding this discrepancy.

Returning once more to visual lexical decision, several researchers have sought to further the understanding of the ambiguity effects reported in this task by identifying their neural underpinnings. For example, Beretta et al. (2005) conducted a replication of the second visual lexical decision experiment reported by Rodd et al. (2002) while simultaneously recording correlates of the neural activity evoked by the task using magnetoencephalography (MEG). This allowed for both the assay of which aspects of neural processing were predictive of performance on the behavioral task (if any, Pylkkänen, Stringfellow, & Marantz, 2002), and where the generators of this activity were located in the brain. In their behavioral results, the latency data were slightly slower overall than those reported in the Rodd et al. study, and the Beretta et al. study reported both a homonymy disadvantage and a polysemy advantage. The MEG results focused primarily

on detecting differences across the word classes in the M350 component. Studying this component was a particularly valuable contribution to the semantic ambiguity literature because of the long history of the M350 and the N400—the M350's approximate EEG counterpart—being relatively insensitive to later processes such as decision-making (Pylkkänen et al., 2002). Consequently, detecting significant differences between the word classes on different measures of the M350 component would be strongly suggestive that it is semantic processing that underlies the behavioral effects, at least in normal controls (but see Taler, Klepousniotou, & Phillips, 2009, for a decoupling of N400 effects and lexical decision latencies in individuals with mild cognitive impairment). Beretta and colleagues reported significantly earlier peaks for words with many senses, and significantly later peaks for words with many meanings. No significant effects were reported in the M350 amplitudes, although there were marginal trends towards larger deflections for words with many meanings relative to words with only one meaning. These results are consistent with the behavioral results, which suggests that the processing that occurs when the M350 is observed is being used as a basis for making the lexical decisions. Further additional support for this possibility arises from the lack of distinction between the conditions on earlier components measuring more surface-level characteristics of word representations (M170, M250), although the authors acknowledge that their methodology was not ideally suited for detecting effects on these components (see also the similar lack of effects on earlier components reported by Taler et al., 2009).

Taken at face value, the Beretta et al. (2005) study replicates the original polysemy advantage reported in Rodd et al. (2002) and suggests that the homonymy disadvantage is a real phenomenon as well. Additionally, this work identifies a neural basis for these effects that is consistent with the brain region that subserves semantic memory. Nevertheless, the inconsistency in the effects of homonymy between the Beretta et al. (2005) study and the Rodd et al. (2002) study are concerning, particularly given that the latter study failed to find a significant effect of homonymy despite including data from twice as many participants. Further inspection of the

methods employed by Beretta and colleagues revealed two potential biases in their methods that may have contributed to a greater likelihood of observing a homonymy disadvantage. First, they made the *a priori* decision to filter out several participants who did not show a clearly identifiable M350 component, which is thought to correlate with semantic access (Federmeier & Laszlo, 2009; Kutas & Hillyard, 1980; Laszlo & Federmeier, 2008). This could have created a selection bias to include only participants with stronger semantic activations and who might therefore show stronger effects of semantic variables such as ambiguity when performing the lexical decision task. Second, the authors failed to include covariates such as frequency and familiarity in their by-item analyses. Failure to include these covariates is later shown to substantially reduce the effect of homonymy observed with these stimuli in a highly-similar study that employed the same word items (see Chapter 5). Consequently, a more conservative interpretation of the results reported by Beretta and colleagues is that lexical decision only consistently produces a polysemy advantage and that this processing advantage is driven by the activation of the semantic representation of the target words.

Other work has focused on gaining a more detailed understanding of the effects of homonymy and polysemy and of the continuous transition between these two artificial categories, as well as on the role of meaning dominance in determining performance. This work is valuable both because it quantifies how the relatedness of a word's interpretations influences processing, and because in doing so it may reveal additional confounds that could explain some of the discrepant lexical decision findings. Important work on this frontier was conducted by Klepousniotou and Baum (2007), who examined the effects of meaning dominance (balanced vs. unbalanced meaning frequencies) on homonymy effects, and of increased sense overlap on polysemy effects; the latter was accomplished by leveraging the linguistic distinction between metaphoric (e.g., <celestial>/<movie> STAR) and metonymic (e.g., <animal>/<meat> RABBIT) polysemy. Whereas metonymic polysemy tends to be linked to a high degree of sense overlap following a relatively regularized distinction between senses, the overlap is reduced and often idiosyncratic in the case

of metaphoric polysemy. In a visual and an auditory lexical decision task, participants responded more rapidly to metonymically polysemous items relative to items in all other conditions and also produced faster responses for metaphorically polysemous items in the auditory task—results that parallel those obtained in N400 amplitudes in healthy aged participants (Taler et al., 2009), which further suggests a semantic basis for these effects.[3] No other differences were observed amongst the conditions. Although it does not appear that item level co-variates were included in any of the analyses conducted by Klepousniotou and Baum (2007), taken cautiously at face value their results lend further support to the existence of a polysemy advantage in lexical decision, for metonymically polysemous items at least, that is independent of the modality in which stimuli are presented.

The state of affairs is slightly more involved for the homonymy effects. The statistical methods employed by Klepousniotou and Baum (2007) were similar to those reported by Beretta et al. (2005) in that they did not include important covariates. This may have artificially inflated the effects of homonymy; however, even employing this potentially more powerful and error-prone test, Klepousniotou and Baum failed to observe a homonymy disadvantage similar to that observed by Beretta and colleagues in their visual task. This finding is also consistent with the null effect of homonymy reported by Rodd et al. (2002) in their visual lexical decision experiment and further supports the hypothesis that in reality, only a polysemy advantage exists. The lack of a homonymy disadvantage in the auditory task does, however, contradict the results of other auditory lexical decision tasks with similar overall latencies. For instance, as already discussed, Rodd et al. (2002) found a significant homonymy disadvantage that was numerically stronger than the polysemy advanage in an auditory lexical decision task that was quite similar to that reported by Klepousniotou and Baum.

One potential explanation for the somewhat inconsistent observation of a homonymy disadvantage in auditory lexical decision may be found in a similar study conducted by Mirman

[3]Taler et al. (2009) report significant differences in waveform amplitudes in the 250-300 ms window, but attribute these to early N400 effects (see also Laszlo & Federmeier, 2011).

et al. (2010). They examined whether the effects of homonymy for balanced homonyms were modulated based on whether homonyms have either noun-noun or noun-verb meanings (e.g., <card>/<boat> DECK vs. <dog>/<tree> BARK). Given that noun-verb homonyms do not share a grammatical class and the noun interpretations often do not have the 'action' representations typically associated with verbs (Watson, 2009), it seemed likely that these items would generate the most competition amongst non-overlapping, inconsistent semantic representations and lead to the most substantial processing disadvantage. Surprisingly, this was not the case in the Mirman et al. study: noun-noun homonyms were significantly slower relative to unambiguous controls, whereas responses to the noun-verb homonyms were not significantly slower than those to the unambiguous items in the item analyses (albeit, which did not include any co-variates). Similar results were also obtained in a four-choice picture-word matching task in which participants were presented with a spoken word and eye-movement latencies to a target image related to the meaning of the word, along with three other distractor pictures, were recorded. This experiment additionally showed that eye-movements to the target image for the noun-verb homonyms were both significantly faster than the noun-noun homonyms and significantly slower than to the unambiguous controls.

Based on the results reported by Mirman et al. (2010), the disagreement in the outcomes of the Klepousniotou and Baum (2007), Mirman et al., and Rodd et al. (2002) studies might therefore be due to whether the homonyms in a given experiment were associated with noun and/or verb meanings. An inspection of the balanced homonyms used by Klepousniotou and Baum and the homonyms used by Rodd and colleagues indicates that most of the items in these studies had both noun and verb interpretations—of the 18 items in the balanced homonym condition in the study conducted by Klepousniotou and Baum, 10 had both noun and verb interpretations in the Wordsmyth dictionary (Parks et al., 1998). All of the 23 items in the auditory lexical decision task reported by Rodd and colleagues, many of which were unbalanced but still resulted in a significant homonymy disadvantage, also had at least one verb interpretation. Of course,

these studies and the re-analyses did not assess whether other factors that may be correlated with noun-noun vs. noun-verb homonymy such as a more sensitive measure of relatedness of meanings, which could also be the cause of these effects (see Appendix B). Nevertheless, taken at face value, these results indicate that if noun-verb homonyms lead to less competition, the weaker and often unreliable homonymy disadvantage observed in these other auditory lexical decision studies might, therefore, be the expected outcome. Similar reasoning would also apply to visual lexical decision tasks as well, although the very weak trends towards a homonymy disadvantage in many of these tasks suggest that this factor alone may not be sufficient to explain the lack of homonymy effects.

Taken together, the individual lexical decision studies have, in some cases, disagreed with respect to the relative differences between homonymous, polysemous, and unambiguous items, but with little exception consistently show at least a rank ordering of these word classes such that polysemous items are responded to most quickly, followed by the unambiguous controls, and lastly by the homonymous items. Careful re-interpretation of many of these studies has also lead to the preliminary conclusion that visual lexical decision is associated with a polysemy advantage in the absence of any effect of homonymy and auditory lexical decision is associated with both a polysemy advantage and a homonymy disadvantage.

However, a series of influential studies conducted by Hino and colleagues (2006, 2010) have been purported to cast doubt on even this basic pattern of results. All of this work was conducted in Japanese, which offered some unique possibilities in terms of nonword foil manipulations in examining ambiguity effects, but which also raises questions regarding the generalizability of these results to single-script languages such as English. In particular, Japanese is composed of two scripts: Katakana, which is analogous to the alphabetic/phonetic representational structure of written English, and Kanji, in which individual characters each correspond to separate morphemes (and to a first approximation, syllables). In one visual lexical decision study, Hino et al. (2006; Experiment 1) reported faster performance for words with less related interpretations

(homonyms) and words with more related interpretations (polysemes) than for unambiguous controls—that is, an overall ambiguity advantage. A similar overall ambiguity advantage has also been reported by Pexman and Lupker (1999) and by Hargreaves et al. (2011), although in neither of those studies were the results significant by items.[4] Hino et al. (2010) replicated and extended the original work by Hino et al. (2006). Once more, they observed an overall ambiguity advantage in a very similar task to that used by Hino et al. (2006; Experiment 1) using Katakana nonwords. They then added Kanji words and nonwords as fillers to their item set in an effort to alter the difficulty of the task (Experiment 2). Kanji words consisted of known pairings of semantically related Kanji characters and Kanji nonwords consisted of unknown pairings of semantically unrelated characters. The addition of these fillers altered the pattern of results substantially, such that only a polysemy advantage was observed. Finally, they re-introduced the overall ambiguity advantage by replacing the Kanji nonwords from the second expeirment with Kanji nonwords formed by transposing the characters in a known pairing of semantically related Kanji characters into an unfamiliar order (Experiment 3).

Taken together, these results apparently contradict those of several similar studies that have been reviewed previously. Several important distinctions between the two experiments may, however, be responsible for these differences in addition to the unknown contributions from conducting these studies in Japanese using multiple scripts. First, the overall ambiguity advantage reported in several of these experiments may be due to how the different word classes have been labeled and contrasted by different groups of experimenters. For instance, in the studies conducted by Hino et al. (2006, 2010), the unambiguous items had fewer interpretations than both the homonymous and polysemous items. These items were then compared to polysemes and homonyms that had approximately equal numbers of interpretations (~5) and differed only in

[4]Notwithstanding, the ambiguous items in the Hargreaves et al. (2011) study did show a processing advantage when the corresponding latencies and accuracies for those items were examined in the E-Lexicon database (Balota et al., 2007), although the validity of this finding is questioned by the results reported in Chapter 5 and in Appendix B.

whether these senses were loaded onto a single meaning or distributed across multiple meanings. The homonymous items used by Hino and collegues could thus be better classified as hybrid items that are both homonymous and polysemous given that these items have more senses than the unambiguous controls, and previous studies have shown to typically produce results that fall between the unambiguous controls and the polysemes (Beretta et al., 2005; Rodd et al., 2002). In contrast, Klepousniotou and Baum (2007) defined unambiguous words as having a single meaning, homonyms as having two meanings, and polysemes as having two senses, and reported the standard polysemy advantage and numeric (but statistically non-significant) homonymy disadvantage that has been present in the other studies reviewed thus far. Nevertheless, even adopting this new nomenclature does not bring the full set of effects reported across the different nonword filler conditions in line with those reported in past lexical decision studies. Those studies would suggest either an increase in the homonymy effect or a decrease in the hybrid effect as a function of increased overall response latency, whereas the Hino et al. (2010) study suggests a non-monotonic homonymy (or alternatively, hybrid) effect as latencies increase as a function of different types of nonword foils. Nevertheless, to date this remains the only study to have reported such a finding. It also employed only 14 target items in each condition, and the authors argued against considering the outcome of item-level statistics in evaluating the significance of the reported effects. Thus, the reliability and generalizability of this finding within Japanese, let alone whether it extends to other languages, is still an open question, and the preliminary summary of the effects of ambiguity in lexical decision that were outlined above still appear to be the most accurate and reliable summary of the existing lexical decision literature.

In addition to these numerous lexical decision studies, important insights into the nature of many semantic ambiguity effects have also been gained using a number of tasks that explicitly required that participants base their responses on the semantic characteristics of a word. The semantic code is inarguably a necessary factor in responding in these tasks, which may provide for a purer assessment of semantic processing and reduce the degree to which the results of

the experiment may be due to other factors that contribute to driving the response system (e.g., orthography, in lexical decision). By requiring the access of particular interpretations of a word to make a response, these tasks can also modulate the semantic representation that must be activated prior to making a response.

In contrast to some of the lexical decision results reported earlier, many previous studies that employed semantic categorization tasks and relatedness judgement tasks reported a significant disadvantage for homonyms (e.g., Gottlob, Goldinger, Stone, & Van Orden, 1999; Mirman et al., 2010; Pexman, Hino, & Lupker, 2004; Piercey & Joordens, 2000). However, Pexman et al. (2004) argued that these results may be due to response competition as opposed to semantic competition because in these tasks each meaning of a homonym supported a competing response (e.g., in a relatedness judgment task RIVER is related to the <edge of water> meaning of BANK, but not to the <money> meaning). To support this claim, Pexman and colleagues varied whether all or only one of the meanings of a homonym supported a particular response and found that the homonymy disadvantage disappeared in the former case. Consequently, these putative semantic ambiguity effects should be more accurately conceputalized as response competition effects and not due to semantics per se. Somewhat paradoxically then, given the success of the Beretta et al. (2005) study at showing a semantic basis for the ambiguity effects in lexical decision, in this case employing a "purer" behavioral assessment of semantics served instead to highlight the potential contributions of the response system in generating some types of ambiguity effects.

Later work by Hino et al. (2006; see also Hino, Lupker, & Pexman, 2002) has, however, shown that attributing the homonymy disadvantage to response competition does not always account for the presence of such a homonymy disadvantage (cf. Pexman et al., 2004). Hino et al. (2006) conducted a series of semantic categorization experiments using the same set of visually-presented Japanese word stimuli for which an overall ambiguity advantage was reported relative to unambiguous words in lexical decision (discussed previously). In different experiments, participants had to decide whether a word refered to 'a living thing,' 'a vegetable,' 'an

animal or a vegetable,' or 'a human position, occupation, or group.' In contrast to much of the previously-referenced work, the interpretations of the ambiguous words never fell into any of these categories and hence consitituted the negative categorization responses. They found that when broad categories were used (i.e., living thing, human), there was a significant processing disadvantage for homonyms (perhaps better labeled as hybrid items, as noted earlier), and that overall latencies were several hundred milliseconds slower in these tasks than in visual lexical decision. However, these results contrasted quite starkly with those of the narrower vegetable and combined animal and vegetable categorization task results, which were also discovered to involve a more extreme separation between the two categories of responses. In those experiments, no effect of ambiguity was observed at all (see also Armstrong, 2007; Hargreaves et al., 2011; Pexman & Lupker, 1999; Siakaluk, Pexman, Sears, & Owen, 2007). Responses were also surprisingly fast—faster in fact than the lexical decision responses for the same word items[5]. Interestingly, similar null or substantially reduced effects of ambiguity have also been reported in lexical decision studies when relatively easy nonwords were employed, although the details of these tasks were often not reported (Azuma & Van Orden, 1997; Borowsky & Masson, 1996; Rodd et al., 2002). Consequently, these results indicate that semantic categorization tasks that involve broad categories and that are associated with slow responses produce a homonymy disadvantage. In contrast, similar tasks that involve narrow categories and that are associated with fast responses do not produce any ambiguity effects, consistent with similarly rapid lexical decision tasks. This latter observation also suggests that at least a portion of these findings are primarily attributable to the representation and processing dynamics that drive the response system at a particular point in time, rather than the specific task that participants are completing.

Relevant to the notion of time as an organizing principle for understanding the discrepancies

[5]This was not the case in Hargreaves et al. (2011), who found some evidence for an ambiguity advantage in a lexical decision task but none in a semantic categorization task with approximately equal latencies. Their results were somewhat atypical, however, because responses in both tasks were unusually slow relative to other similar studies (e.g., Hino et al., 2006).

amongst ambiguity effects, Siakaluk et al. (2007) also found that semantic categorization tasks that involved narrow categories such as 'animal' did show homonymy disadvantages when a slower go/no-go version of semantic categorization was used instead of the standard yes/no version, which does not show such effects. Several theories and supporting models of the decision system and lexical-semantic system underlying responses in each of these tasks suggest that the underlying mechanics of these systems are highly similar in both tasks. This in turn suggests that it is primarily increased processing time, as opposed to a qualitative task difference, that underlies the different effects that are produced when the nature of the response is manipulated, holding all else constant (Gomez, Ratcliff, & Perea, 2007; Siakaluk, Buchanan, & Westbury, 2003; see also supporting behavioral evidence from Perea, Rosa, & Gómez, 2002).

There has also been some recent work exploring the neural basis of these different semantic categorization effects. Hargreaves et al. (2011) found that there was significantly greater activation in the left inferior frontal gyrus—which in their view, may provide domain-general inhibition when representations from, for instance, as ambiguous words compete to become activated—for the ambiguous items in a narrow 'animal' category, even in the absence of significant behavioral results. This would imply that on the time-scale of fMRI, which may reflect additional processing that occurs post-response, both meanings are being activated and frontal regions are recruited to try to adjudicate amongst responses. This latter result is not without some controversy, however, and contrasts with the account of Rodd, Johnsrude, and Davis (2010) regarding the role of the left frontal gyrus. Using a behavioral task-interference paradigm, they concluded instead that the left frontal gyrus becomes active only several seconds after the ambiguous stimulus and contributes to the re-activation of a previously de-activated meaning, and not during the initial ambiguity resolution process that would be contributing to a behavioral response on the order of $500 - 1000$ ms. Given the paucity of studies that have investigated the neural underpinnings of these ambiguity effects, the inconsistencies in these results preclude drawing strong, specific conclusions from either of these studies. Nevertheless, these studies are in agreement, in prin-

ciple if not in the details, that the frontal lobes playing a role in the later processing dynamics associated with tasks that assess ambiguous word comprehension.

Summary. Factoring in issues of statistical control and nomenclature differences, as well as the consistency of particular findings across different experiments, the most reliable semantic ambiguity effects reported in the previous section are as follows: Visual lexical decision tasks tend to show an early polysemy advantage and at best a very weak homonymy disadvantage. Auditory lexical decision tasks produce similar results but with more reliable homonymy effects and with considerably longer overall latencies. Semantic categorization and relatedness judgment tasks that involve broad categories and that avoid response competition confounds show considerably longer latencies than comparable visual lexical decision tasks (albeit slightly shorter latencies than auditory lexical decision); these tasks produce a homonymy disadvantage. Similar tasks that involve narrow categories produce very fast responses and, as in similarly very rapid lexical decision tasks, produce no ambiguity effects whatsoever. The neuroimaging data—broadly construed to include the EEG/MEG results—further support a semantic basis for many of these effects and not one based on earlier perceptual/orthographic processing or later response selection processes. These results also suggest that the time-course and nature of the temporal lobe and frontal lobe contributions to processing are at least partially non-overlapping, with the frontal lobes conributing to the later suppression of an inappropriately-activated meaning of a word.

Studies that involve context

The results of studies that involve context that biases the interpretation of an ambiguous word, which are discussed in detail in the following section, provide addititional insights into how ambiguous words are processed and how context constrains the activation of a particular interpretation. Key studies that fall under this rubric are described below. Note that the exact nature of the 'context' provided in many of these studies varies considerably, from individual words to full

sentences, and there is as-yet no comprehensive agreement with regards to what exactly makes a context biasing, although there is agreement in the subjective assessments of what constitutes a biasing context (Paul, Kellas, Martin, & Clark, 1992). Importantly, however, the findings that have been reported are similar when either individual words or full sentences serve as biasing contexts, and these findings cannot all be explained by simple lexical association (Simpson, 1994), contrary to one previous theory (Seidenberg, Tanenhaus, Leiman, & Bienkowski, 1982). Rather, the same ambiguity resolution mechanism appears to operate regardless of the type of context that is employed, as would be expected if the ambiguity is resolved by a general comprehension mechanism (McClelland et al., 1989).

A classic study examining context effects was reported by Swinney (1979; see also Conrad, 1974; Kintsch & Mross, 1985; Oden & Spira, 1983; Onifer & Swinney, 1981; Seidenberg et al., 1982). He presented participants with auditory recordings of sentences containing homonyms (e.g., "Rumor had it that for years the government building had been plagued with problems. The man was not surprised when he found several spiders, roaches, and other *bugs* in the corner of his room"; there was no emphasis added to the homonym in the original study). Between 0 and 1000 ms following the presentation of the homonym, either a nonword or a word was presented visually that prompted a lexical decision. The word items were either consistent (e.g., ANT) inconsistent (e.g., SPY) or unrelated (e.g., SEW) with the context evoked by the sentence. Swinney found that at short latencies all of the meanings of the homonym were primed, whereas at long latencies priming was observed only for the contextually-appropriate meaning. Performance for the inconsistent meaning was not, however, any slower than that for an unrelated word. A similar reduction or elimination of priming for the inconsistent meaning of a homonym after a very brief period of time had elapsed between the homonym and the target word was also obtained in spaced-presentation repetition effects of same versus different meanings in lexical decision and naming (Masson & Freedman, 1990; Van Petten & Kutas, 1987). The original Swinney (1979) study was also followed up by a similar study that focused on polysemes instead of homonyms

(Williams, 1992). The results of this study also showed early activation of both the consistent and inconsistent senses, but in contrast to Swinney (1979), a significant processing advantage was observed even when words related to the inconsistent sense were presented after a longer period of time, although the latencies for the inconsistent senses were numerically slower than the latencies for the consistent senses.

Van Petten and Kutas (1987) conducted a similar experiment in which they recorded ERP waveforms at both short (200 ms) and long (700 ms) stimulus onset asynchronies (SOAs). In most respects, their task paralleled that reported by Swinney (1979) but their participants did not make explicit responses, thus avoiding any potential contamination from the response system in the ERP waveform. They found that N400 amplitudes recapitulated the behavioral effects and could therefore be the neural source that drives these different effects. Specifically, they found that at the long SOA, words related to the inconsistent meaning showed a similar large N400 amplitude as unrelated words. In contrast, at the short SOA the inconsistent items, which showed significantly larger N400s than the consistent items, nevertheless elicited a smaller N400 relative to the unrelated words. These effects were consistent with the original conceptualization of the N400 as indicating of semantic incongruencies, such that context-sensitive interpretation activation is limited during early processing (Kutas & Hillyard, 1980; see Laszlo & Plaut, 2011 for an updated view).

The previous studies establish that early in processing both interpretations of an ambiguous word are partially activated. Later, a strong contextual constraint causes the de-activation of the contextually-inappropriate interpretation (albeit to a lesser extent for polysemes). However, with the exception of the weak gradation of consistent, inconsistent, and unrelated meanings in the N400 data reported by Van Petten and Kutas (1987), the previous results do not clearly indicate whether initial semantic access is necessarily contextually independent (Swinney, 1979), or whether this late effect of context was simply late in these studies because a relatively weak contextual bias was used. More recent work directly addresses this issue. For instance, Klepousniotou

(2002) reported the results of a lexical decision task in which legal nonword foils and target homonymous, metaphorically polysemous, and metonymically polysemous words were first primed by sentences consistent with either the dominant or subordinate interpretation of the ambiguous words. She found that relative to unambiguous controls that were unrelated to the context, all of the target ambiguous words showed a processing advantage from context. This indicates that context exerts a small but significant effect on early processing.[6] Seidenberg et al. (1982) also reported similar early contextual biases, although not quite as early as in Klepousniotou and Baum's study. Additionally, the magnitude of the advantage increased as semantic overlap between the interpretations increased, paralleling the processing advantage as a function of semantic overlap observed in later work (Klepousniotou & Baum, 2007). Together, these results indicate that context can have an early effect on processing, but that context-independent processing tends to dominate the early semantic ambiguity effects.

To further explore this possibility, several other studies have also examined the effects of extremely biasing context on processing. These studies have reported that the inconsistent meaning of a homonym may be not be activated at all after the first 120 ms of processing (Martin, Vu, Kellas, & Metcalf, 1999; Simpson, 1981; Tabossi, 1988)—considerably earlier than the latencies at which Swinney (1979) reported context-sensitive processing, albeit not so early as to be observed immediately after the presentation of the word (Van Petten & Kutas, 1987). This effect has been more consistently demonstrated for the suppression of a subordinate meaning in a dominant context than when the context biases the subordinate interpretation and a prime related to the dominant interpretation is presented (Forster & Bednall, 1976; Hogaboam & Perfetti, 1975; Neill, Hilliard, & Cooper, 1988; see also the explicit ambiguity judgment tasks reported

[6]Surprisingly, Klepousniotou (2002) also reported faster responses for homonyms and metaphoric polysemes relative to metonymic polysemes in her ambiguous control words, in-line with the ambiguity advantage reported by Hino et al. (2006, 2010). However, the experimental items do not appear to have been very well controlled, which casts doubt on the validity of this inference. A later isolated-word lexical decision task that employed a better-controlled set of items also failed to replicate these results (Klepousniotou & Baum, 2007).

by Simpson, 1981). This may be because unless the prime is extremely strong, context may have the effect of transiently rendering the subordinate meaning to be of equal or slightly greater apparent frequency than that of the dominant meaning. In that situation the prediction is that both meanings should be accessed. These results provide further support for the graded nature of contextual influences and discount earlier all-or-none views of early contextual constraint (Swinney, 1979), but are still consistent with the notion of very early processing being context-free, at least to a first approximation.

The lack of selective inhibition for the contextually-inconsistent meaning of a homonym in the studies just reviewed additionally supports the proposed role of inhibition as a non-specific activation regulation mechanism (Hargreaves et al., 2011). There have, however, been some recent data that question this simplified view of inhibition. Nievas and Marí-Beffa (2002) proposed that the task demands and degree of contextual bias in many of these experiments may have under-powered the detection of any selective inhibition effects. To support their claims, they conducted a modified version of the standard lexical decision task. In this task, participants were first presented with two primes, one a homograph (i.e., a word for which the printed form was associated with multiple meanings, but for which each of those meanings could potentially be associated with different auditory forms, e.g., <storm>/<turn> WIND) and the other an associate related to one of the meanings of the homograph (e.g., TURN-WIND). They then completed a lexical decision task wherein a probe word item was either be related or unrelated to one of the meanings of the homograph in the form of longer latencies for those items. In the case of a related item, it could further be related to either the consistent or inconsistent meaning activated by the primes. As expected based on previous studies, they found that when the word probe was related to the consistent context, there was a large decrease in latencies. However, they also found evidence for selective inhibition of probes related to the inconsistent meaning of the homograph. More detailed analyses of participants expressing different speed-accuracy trade-offs showed that this selective inhibition effect increased as more time passed—average-

to-fast responders did not show a substantial inhibitory effect, but the slowest responders with the highest accuracy did. They interpreted these results as evidence for selective inhibition, likely resulting from inhibitory control. A similar conclusion was also drawn by Reder (1983), who reported negative priming for the contextuall- inconsistent interpretation of a word, as assessed by interpretation errors. These conclusions are consistent with theories of negative priming in which inconsistent items, broadly construed, are inhibited below baseline (Gernsbacher & Faust, 1991; Houghton & Tipper, 1996; Tipper & Driver, 1988). This result also suggests a more sophisticated role for inhibition in explaining the details of some ambiguity effects. Nevertheless, the effects of selective inhibition reported by these studies are generally quite small relative to the facilitation effects observed in other studies.

Several developmental studies have also employed methods simliar to those used in many of the previously-discussed lexical decision studies that involved contxtual priming. These studies offer unique insight into ambiguous word comprehension and how these processes change over the course of development (Booth, Harasaki, & Burman, 2006; Khanna & Boland, 2010; Simpson & Foster, 1986; Simpson, Lorsbach, & Whitehouse, 1983). The results of these studies indicate that early on in development, children (age 5-7 years) initially activate all of the meanings of known homonyms regardless of the context in which the words are encountered (Khanna & Boland, 2010). The presumed development of inhibitory processes to suppress the contextually-inappropriate meaning is slower and is not largely completed to near-adult levels until much later (9-12 years). The results of these studies also show that the later/slower development of inhibition leads to weaker inhibitory effects in the earlier age groups. Consequently, contextually inappropriate meanings that are usually suppressed by adults in the first 500 ms of processing may take closer to 1000 ms for children to suppress, if they are suppressed at all (Khanna & Boland, 2010).

Similar tests have also been completed with Broca's aphasics (Swaab, Brown, & Hagoort, 1998) and patients with Alzheimer's dementia (Balota & Duchek, 1991; Faust, Balota, Duchek,

Gernsbacher, & Smith, 1997). These populations typically show slower and weaker inhibitory effects and can shed additional light on the role of inhibition in a the typical system. The results of these studies identified similar patterns of reduced inhibition to those observed in early development. They also showed additional enhancement of a contextually-appropriate interpretation consistent with an under-performing inhibitory mechanism. Reduced inhibition has also been reported to a lesser extent in older controls as well (Dagerman, MacDonald, & Harm, 2006). Nevertheless, the results of these studies should be interpreted cautiously because evidence from other tasks such as relatedness judgments calls into question whether children (and presumably, adults with neurological impairments) attend to the context properly (Khanna & Boland, 2010), or whether limited working memory capacity may restrict the influence of context (Khanna & Boland, 2010, but see Mason & Just, 2007, for evidence that increased working memory increases the degree to which multiple meanings are activated).

Thus far in the review, a series of semantic ambiguity effects has emerged in a variety of tasks. An important question to ask, however, is whether such ambiguity effects are observed in a more naturalistic setting and are not caused by task demands and how they interact with other systems (e.g., as illustrated by the semantic categorization study conducted by Pexman et al., 2004). Evidence supporting the semantic basis for the effects of context on homonym processing has been reported by Frazier and Rayner (1990) using a more naturalistic reading paradigm and a measure of fixation duration obtained via eye-tracking (Duffy, Morris, & Rayner, 1988; see Frisson & Pickering, 1999, for a follow-up focusing on different types of polysemes). They examined participants' gaze durations for homonyms and polysemes relative to unambiguous controls presented either before or after a biasing context. They found that subsequent to the presentation of a biasing context, participants' gaze durations were longer for all ambiguous words relative to the unambiguous controls. This suggests that resolving the semantic ambiguity impedes processing to some extent for all ambiguous words. Similar results were reported by Piercey and Joordens (2000), who observed a processing disadvantage for polysemes when contextual integration must

take place to generate a response.[7] However, when disambiguating context was not presented before the ambiguous words, responses were slower for the homonyms only, suggesting that the presence of multiple meanings in and of itself is sufficient to generate a processing disadvantage. Theories of eye-movement initiation (e.g., Reichle, Pollatsek, Fisher, & Rayner, 1998) also posit that eye-movements are essentially initiated once the bulk (although not necessary all) of the processing of the perceived word has been completed (Reichle et al., 1998), which suggests that this processing disadvantage is present near the point at which the maximal amount of processing of a word has been completed given the available contextual information. This possibility is particularly interesting because the presence of such a homonymy disadvantage in the absence of a strongly biasing context was also observed in the semantic categorization tasks involving broad categories reported in the earlier portion of the review. A parsimonious view of these data would therefore be that both of these tasks activate and are influenced by the same semantic representation.

The extremely short fixation durations in the Frazier and Rayner (1990) data, which were typically less than 300 ms, do, however, raise concerns about the similarity between these results and those from other tasks that show similar results at much longer latencies when button presses instead of eye movements are recorded. Clearly, button-presses in behavioral tasks may simply take longer to execute because of the time needed to select a response and to physically press the button and account for a large part of the ~500 ms discrepancy between the eyetracking data and behavioral datas. This is supported by the fairly comparable (although slightly shorter) latency discrepancy between the behavioral results and the MEG results reported by Beretta et al. (2005). Direct neurophysiological measures may therefore be a more appropriate comparison to the eyetracking results (Sereno, Rayner, & Posner, 1998). In the N400 latency analyses, clear

[7]The original Piercey and Joordens (2000) study reported this effect as a homonymy disadvantage. However, the items used in that experiment were taken from the Borowsky and Masson (1996) study that Rodd et al. (2002) found to be primarily polysemous and not homonymous. The labeling scheme used in the Rodd et al. study is therefore adopted here.

semantic ambiguity effects can typically be detected on the order of 250 ms and sometimes slightly earlier (Laszlo & Federmeier, 2011; Taler et al., 2009; Van Petten & Kutas, 1987). Consequently, in principle the results from each of these different tasks and types of measures may all be driven by activation in lexical-semantic memory.

In addition to focusing on how ambiguous words are processed, some researchers have also investigated the representations of ambiguous words themselves and how they interact with comprehension processes. One series of such studies was conducted by Klein and Murphy (2001, 2002), who aimed to determine how specific senses of polysemes are represented in memory. Specifically, their studies focused on determining whether multiple related senses overlap and share a substantial 'core' interpretation, as had been assumed by some other researchers (e.g., Frazier & Rayner, 1990; Rodd et al., 2002), or whether each sense was represented essentially separately as is generally assumed for the different meanings of homonyms. Klein and Murphy (2001) studied these issues by presenting a polyseme twice in two phrases that tapped either the same sense or a different sense of the word (e.g., PAPER as a material or a publication printed on that material; see also Bainbridge, Lewandowsky, & Kirsner, 1993). They found that presenting a second sentence using the same sense of the polyseme increased memory for the polyseme in a later memory test and also decreased latencies in a sensicality judgement task relative to sentences involving different or neutral/non-sensible senses (e.g., SHREDDED PAPER could be used to prime WRAPPING PAPER [sensible / same sense], HISTORY PAPER [sensible / different sense] or SPACE PAPER [non-sensible, because this does not convey any consistent, coherent interpetation]). Moreover, the latency differences between the same- and different-prime conditions in the sensicality judgment were only slightly smaller for polysemes than for homonyms, and polysemes showed facilitation for consistent senses and inhibition for inconsistent senses relative to a neutral baseline (see also the similar effects for homonyms only that was reported by Seidenberg et al., 1982). Similar results were obtained in Klein and Murphy's (2002) follow-up to these original findings. They reported that phrases using different

senses of a polyseme were only rarely categorized together in the presence of a thematic lure in a forced-classification task (e.g., participants preferred to group SMOOTH CLOTH and WRAPPING PAPER together rather than WRAPPING PAPER and LIBERAL PAPER), and the low rates of grouping the polyseme senses together were only slightly (but significantly) higher than those of grouping the different meanings of the homonyms together. Taken together, Klein and Murphy (2001, 2002) argued that these results indicate that the semantic overlap across related senses is minimal. Consequently, relatedness must be encoded separately from semantic overlap (see also the lexical decision data from Hino et al., 2010; Neely, 1991).

Despite the number and variety of studies that are consistent with Klein and Murphy's (2001) conclusion regarding the representation of relatedness, there is nevertheless reason to be skeptical of this conclusion. First, in an extension of Klein and Murphy's (2002) work, Pylkkänen, Llinás, and Murphy (2006) replicated the behavioral task employed in the original study while recording MEG data. In contrast to the original behavioral findings, they found significantly earlier M350 latencies for polysemes and delayed M350 latencies for homonyms (particularly in the left hemisphere). These results suggest that Klein and Murphy's behavioral task and associated measures may not be particularly sensitive to the actual semantic processing that occurs during these tasks (see also Beretta et al., 2005). They also suggest that each sense of a polyseme may share a core interpretation and need not be represented separately, as is assumed to be the case for homonyms.

Klepousniotou et al. (2008; see also Williams, 1992) also failed to reproduce the same patterns of effects reported by Klein and Murphy (2001) when using a finer-grained break-down of the relatedness continuum into low-, medium-, and high-overlap items roughly equivalent to homonyms, metaphoric polysemes, and metonymic polysemes instead of employing only polysemes that were "clearly distinct in meaning" (Klein & Murphy, 2002, p. 568). In contrast to Klein and Muphy's (2001, 2002) results, they found that the latency difference across phrases priming two different interpretations of an ambiguous word were relatively small for

the high-overlap items. Similar to Klein and Murphy's results, they did, however, find a large latency difference for the medium- and low-overlap items, which were non-significantly different from one another. Regrettably, their unambiguous control condition was not successful at establishing baseline performance (see Klein & Murphy, 2001; Klepousniotou et al., 2008; Plaut & Booth, 2000, for discussion regarding the establishment of appropriate baselines in priming experiments). Whether all of these effects should be viewed exclusively as competitive slowdowns, or whether a portion of these results should be viewed as enhanced processing relative to a neutral control, is consequently unclear. Nevertheless, these results are compatible with there being substantial featural overlap in the case of the high-overlap items, but that this overlap and/or the processing advantage associated with high overlap decreases more rapidly between the high- and medium-overlap items relative to the medium- and low-overlap items. Consistent with this proposal, Klepousniotou et al. (2008) found that most of the polysemous items used by Klein and Murphy would be considered to have either low or medium semantic overlap according to their classification scheme, and which were also not significantly different in their study. This explanation has not gone without some controversy, however, and Hino et al. (2010) suggested that when a word's senses/meanings are highly overlapping as in Klepousniotou et al. (2008), they may actually be better thought of as being unambiguous (p. 182). However, metonymic polysemes such as those used by Klepousniotou et al. (2008) have previously been shown to produce different effects than unambiguous words with only a single sense (Klepousniotou & Baum, 2007), which invalidates this claim.

Summary. Having identified the likely causes for many of the disagreements reported in this literature, the data from the studies that involved biasing context can now provide important additional insight into how ambiguous words are processed. Early on, contextual effects are typically weak and multiple interpretations may be activated unless an extremely biasing context is employed. As the influence of context increases, the contextually-appropriate interpretation remains active but the inappropriate interpretation's activity decreases as a function of its relatedness to

the appropriate interpretation. This decrease generally appears to drop the activation of the inappropriate interpretation to the same level, or slightly below, that established by neutral control words; the effects of selective inhibition are, however, are quite weak at best. Developmental and neuroimaging data also provide support for these effects being caused by processing dynamics in the anterior temporal lobe regions that have a long history of being associated with semantic processing (for an overview, see McClelland & Rogers, 2003), and for the separation—both functional and neuroanatomical—of some inhibitory processes from the main semantic hub.

2.2 Theoretical Accounts of Semantic Ambiguity Effects

Having established the key empirical phenomena that are reliably reported in the semantic ambiguity literature, as re-emphasized in the prevous summaries, the next section reviews several theoretical accounts aimed at capturing subsets of these data. This review serves as the basis both for introducing these accounts, which play an important role in subsequent discussions throughout the dissertation, and for identifying the strengths and weaknesses of past approaches that should be addressed in developing an improved account. Given the large number of alternative accounts of specific phenomena that have been presented over the last 40 years, this is not intended to be an exhaustive review of all past theories. Rather, the review focuses on the somewhat broad accounts of substantial numbers of ambiguity effects. Accounts that were explicitly described as task-specific and that are not relevant to developing a general theory of how ambiguity is processed and represented are not included (e.g., Van Petten & Kutas, 1987), with the exception of the lexical decision task, given how prominently it has been featured in the extant literature. The different accounts can basically be grouped into whether the locus of ambiguity effects is believed to be in the orthographic,[8] semantic, or response selection systems. Accounts

[8]Presumably, this account would incorporate both orthographic and phonological surface forms depending on the modality of the task. However, these accounts are labeled as 'orthographic' here because most past research has focused on visual word recognition.

falling in each of these categories are reviewed in turn.

Orthographic accounts. There is a long history of theoretical accounts of lexical decision performance being either primarily (Balota & Chumbley, 1984; Coltheart et al., 2000; Grainger & Jacobs, 1996; Hino & Lupker, 1996; Hino et al., 2002; Kawamoto, 1993; Kawamoto, Farrar, & Kello, 1994; Pexman & Lupker, 1999), or at least partially (Seidenberg & McClelland, 1989), due to the processing of the orthographic as opposed to the semantic representation of a word. For instance, Kawamoto (1993; see also Kawamoto et al., 1994) reported a connectionist simulation of how the time-course of orthographic feature activation (hereafter refered to as settling) could account for the overall ambiguity advantage that was the accepted pattern of results at that time. According to this account, ambiguous words received less consistent feedback from semantics because different semantic features are activated for the different meanings. As a result, ambiguous words develop stronger mappings within orthography, whereas unambiguous words relied more on the consistent feedback from semantics and developed weaker intra-orthographic mappings. Building on a similar theme, Hino and Lupker (1996; see also Hino et al., 2010, 2002, 2006; Pexman et al., 2004; Pexman & Lupker, 1999) suggested that when an ambiguous word engages semantics it will elicit more semantic activation than an unambiguous word. This activation will then feed back and more strongly activate the ambiguous word's orthography, leading to an ambiguity advantage. Support for this prediction is also provided in the form of slower lexical decision latencies for words that have high frequency synonyms (Hino et al., 2002). Relatedly, but within the orthographic-to-phonological subsystem and with effects more in line with the patterns of data reviewed presently, Seidenberg and McClelland (1989) reported slower settling for homographs (e.g., <breeze>/<to turn> WIND) than non-homographs because there were weaker orthographic-to-phonological mappings for the homographs because of the reduced consistency of these mappings. Similar work by R. L. Newman and Joanisse (2011) with homophones also supports this general position in a study of homophone processing, and also points to a later frontal contribution for longer and more complete settling.

There are several problems with these accounts. Obviously, one major problem is that several of these accounts, as they were initially proposed, simply do not fit the updated patterns of ambiguity effects that were reviewed in the previous sections (e.g., the Joordens & Besner, 1994, model predicts a homonymy advantage, whereas a homonymy disadvantage is typically observed in lexical decision). There is also a strong theoretical motivation for not preferring the accounts that involve semantic-to-orthographic feedback: The proposal that lexical decisions would be based on orthography, which is itself indirectly influenced by semantics, is unparsimonious and illogical without substantial independent support. If semantics interacted with orthography and produced different types of activation patterns, why would this activation necessarily need to feed back to orthography to influence performance? Rather, it seems more straightforward for the response system to simply monitor semantics directly rather than only to monitor orthography and need to wait for activation to feed back before responding. Nevertheless, the general proposal that orthography may either directly or indirectly contribute to lexical decisions is still a reasonable intuition in abstract terms. This notwithstanding, an orthographic account, however relevant for capturing the lexical decision data and associated ambiguity effects, is fundamentally incapable of capturing a large portion of the inarguably semantic effects reported in a large number of the reviewed experiments (e.g., the effects reported in semantic categorization tasks).

Semantic accounts. Several semantically-based accounts of ambiguity effects have been proposed. One influential theory was outlined by Joordens and Besner (1994). They theorized and provided supporting connectionist simulation data for an ambiguity advantage arising because the initial starting location for semantic settling, as determined by the previous word in a lexical decision experiment in which the order of stimulus presentation was random, was on average closer to one interpretation of a homonym than to the single meaning of an unambiguous control (see also Jastrzembski, 1981). This advantage was limited to connectionist networks with small numbers of hidden units and extremely high (74%) error rates, however, and dissappeared in larger networks.

Although of considerable theoretical value for other reasons, for present purposes this work may now best serve as a cautionary note for future simulations—very small simulations may lend themselves to supporting theoretical proposals that are incompatible with the representations and processes at play in the brain, and are best employed in conjunction with larger-scale simulations. Specifically, the Joordens and Besner (1994) account of the ambiguity advantage hinged on the notion that starting from a random point in semantic space, one meaning of a homonym would be substantially closer to one of the semantic representations associated with that word (i.e., a proximity advantage). This would also avoid any potential competitive effects because one meaning already dominates (by virtue of its proximity advantage and increased initial activation). However, from the central limit theorm, it is known that the likelihood of this occuring to a substantial degree decreases as the size of the representational space grows large and both meanings are expected to be at approximately equal distances from the random starting point. Fortunately, this fact alleviates concerns with regard to how connectionist principles could, by virtue of what turned out to be oversimplifying assumptions, lead to the emergence of contraditory patterns of predictions and disagreement with the empirical data that were just reviewed.

Another type of semantic account has focused on reducing the putative semantic ambiguity effects that have been reported to effects of featural richness and the related notion of overall semantic activation. It is well known that featurally rich items, such as concrete nouns (e.g., DOG) are typically processed more rapidly than less rich items such as abstract nouns (e.g., JUSTICE; Plaut & Shallice, 1993). Consequently, if semantically ambiguous items differed in terms of semantic richness, this could be the actual cause of the "ambiguity" effects. Polysemes in particular, which may be able to activate large numbers of core features, have been hypothesized to produce an ambiguity advantage that is in part due to richness (Hargreaves et al., 2011; Klepousniotou & Baum, 2007; Pexman, Hargreaves, Siakaluk, Bodner, & Pope, 2008; Pexman, Lupker, & Hino, 2002; Rodd et al., 2002). However, other studies, including those reported in

later chapters, that control for a measure of richness (imageability) that was made to be orthogonal to the number and relatedness of meanings of the target word still yield significant ambiguity effects (see also Rodd et al., 2002, Experiment 2). Thus, featural richness may be more appropriately treated as a potential confound in ambiguity experiments, as is the case with factors such as frequency and familiarity, rather than as an explanation of all ambiguity effects in its own right.

A third account has focused on the semantic settling dynamics that unfold over time for words with different numbers and relatedness (operationalized in terms of feature overlap) of interpretations to explain the lexical decision results (Rodd et al., 2004). This account was successful in predicting that polysemous words were processed more quickly than unambiguous words, which were in turn processed more quickly than homonymous words. An explicit instantiation of this account using the connectionist framework also successfully substantiated these predictions, although it did predict a stronger homonymy disadvantage and a weaker polysemy advantage, whereas the opposite was true in most of the behavioral data. This account did, however, suffer from the critical limitation of being explicitly cast as a model of semantic processing in lexical decision, and was not designed as a general model of comprehension per se. Consequently, this account cannot be extended to capture the data from other tasks without extensive modification.

Arguably, the semantically-based account that has captured the broadest set of different results was the one proposed by Piercey and Joordens (2000). According to this theory, different tasks required different degrees of precision in the underlying semantic code for participants to generate a fast and accurate response. For instance, tasks such as isolated-word lexical decision require only the activation of an imprecise semantic representation and could essentially be based on whether a given word form was associated with any interpretation(s) whatsoever; the full activation of a particular interpretation as constrained by context was not necessary. In such tasks, ambiguous words would be associated with an ambiguity advantage because more semantic features are active early in processing, before the features of the contextually-inconsistent interpretation are suppressed. In contrast, other tasks require a more precise semantic represen-

tation that is constrained by context. Activating such a precise semantic representation would take more time for ambiguous words because these words would need to interact with context to activate an appropriate interpretation, whereas unambiguous words could begin activating their only correct interpretation without factoring in the constraints provided by context. This theory has many of the qualities of the desired type of account, in that it is based on semantic processing and holds promise for accounting for a range of different effects. However, it suffers from the critical flaw that in its current form it predicts an ambiguity advantage in lexical decision tasks—particularly for homonymous words, whereas if there is any effect of homonymy in lexical decision, it is a processing disadvantage. The data from semantic categorization tasks also pose a similar problem for this account.

Decision system accounts. Hino et al. (2006; see also Hino et al., 2010) proposed that the ambiguity advantage in lexical decision and the homonymy disadvantage in semantic categorization arise not because of settling dynamics in semantics—which they assume are constant and hence are unable to explain different patterns of results—but are due, instead, to qualitative differences in how decisions are made based on either the orthographic or the semantic code. Additionally, the general notion of decision system differences may explain the different patterns of results obtained in semantic catogorization tasks using broad and less distinct (e.g., living thing) or narrow and more distinct categories (e.g., vegetable), as well as performance on relatedness judgment tasks when all or only some of an ambiguous word's interpretations are consistent with a particular response (Pexman et al., 2004)—different types of semantic information monitoring are needed in each of these tasks, and some effects appear to be the result of response competition. Evaluating the validity of these types of accounts is a non-trival undertaking, however, because Hino et al. (2006) did not provide a detailed mechanistic description of the decision system and how it interacts with semantics and orthography. Consequently, the degree to which an explicit system could produce a substantial portion of effects on the basis of independently motivated and verifiable model parameters remains an open and substantial empirical question at best, and is

the motivation for some of the investigations reported later. Nevertheless, such a system appears destined to be unable to account for the more naturalistic and direct assays of comprehension processes (e.g., eyetracking and EEG/MEG data), which appear to be directly linked to semantic processing.

Critical Evaluation of the Accounts and Objectives for an Improved Account

The existing accounts all suffer from serious limitations that rule them out as general accounts of a broad range of ambiguity effects. The main limitations can be grouped into three categories. First, several of the accounts are narrow in the scope of issues that they addressed (e.g., Rodd et al., 2004; Van Petten & Kutas, 1987). The present review excluded accounts that addressed only a single empirical phenomenon with the exception of the lexical decision task, but even an account of lexical decision or of a few key phenomena leaves much to be desired if it cannot be extended readily to other data. Second, many of these accounts are underspecified in that explicit demonstrations of how they operate often have not been proposed (e.g., Hino et al., 2006). Without such demonstrations, many important details related to these accounts are not clear, and it remains questionable whether an explicit implementation would generate the predicted effects. Third, many of these accounts are limited by the simple fact that they produce effects that run contrary to key effects in the literature (e.g., Joordens & Besner, 1994; Kawamoto, 1993), with the strong polysemy advantage and the weak-or non-existent homonymy disadvantage in lexical decision being particularly problematic for most accounts.

A better account should address these limitations. First, it should aim to capture a broader range of effects than those from a single task, although a solid account of lexical decision would be valuable given the literature's strong focus on this task. Second, if the model invokes complex representation and processing assumptions, those assumptions should be directly testable and not left underspecified and unclear. The chapters that follow describe the development, the

substantiation, and the evaluation of an account that aims to satisfy these goals.

Chapter 3

The Settling Dynamics Account

On the basis of the insights gleaned from the literature review, the following chapter introduces a theoretical account of ambiguous word comprehension that captures a wide range of semantic ambiguity effects. The overarching premises of this theory are that the range of semantic ambiguity effects that have been reported are due to two main factors: 1) the temporal dynamics underlying a) co-operation and competition amongst an ambiguous word's semantic representations and b) contextual integration, and 2) the overall precision of the semantic code needed to make a decision in a given task. Together, these dynamics cause various ambiguity effects to emerge at different points in time within a system instantiating a single neurobiologically-inspired computational formalism. The first section introduces the account in more detail, the second section describes how the account captures, in principle, a range of ambiguity effects, and the third section describes key challenges to this account and how they can be addressed via computational and behavioral investigations.

3.1 Introduction to the Settling Dynamics Account

As noted by Simpson (1994), "it has become quite apparent that any explanation of ambiguity processes that does not include an account of changing activation patterns over time simply can-

not hope to capture the complete picture of the processes relevant to the selection of a single appropriate meaning" (p. 365). From the review presented in the previous chapter, and without denying that other factors such as the processing dynamics in the response system may make important contributions to understanding the full details of particular effects, it is clear that for a relatively simple account to capture a wide range of semantic ambiguity effects, the temporal settling dynamics in semantics should be the primary focus of the theory. Given that these dynamics will necessarily be somewhat complex so as to capture a wide range of findings, it is also important to develop the account within the context of a general framework and well-established computational formalism. This will allow for the explicit instantiation of the theory's key principles so as to demonstrate that the hypothesized dynamics are, in fact, produced. It will also allow for the generation of new predictions that can be used to confirm the basic principles of the account and guide future research. Given the connectionist framework's long history of statisfying these requirements and its commitment to the core representation and processing assumptions that are proposed to underly a range of semantic ambiguity effects, this framework was selectd to serve as the backdrop for these efforts. To ensure the generality of the account and avoid overfitting past findings, the key principles of the settling dynamics account are grounded in the general principles thought to underlie word comprehension, the nonlinear processing dynamics and representation assumptions of the connectionist framework, and the semantic representations that the empirical data suggest are associated with ambiguous and unambiguous words.

The basic assumptions and predictions of the account are introduced below. On the basis of the empirical data, the effects of context are relatively weak during early processing, so the account is unpacked in two sections: The first considers a simpler set of dynamics that are theorized to dominate early processing that occurs in the absence of biasing context. This section covers early processing until the semantic representations are driven into what Joordens and Besner (1994) termed a 'blend state', in which all interpretations of an ambiguous word are partially activated. The second section focuses on the later dynamics that involve context. To be clear,

however, this should not be interpreted a a staged theory of context-sensitive word comprehension (e.g., Swinney, 1979); the effects of context are assumed to start off weak and gradually increase in magnitude while the word is being processed. This stage-like division is employed strictly to simplify the description of the account. Additionally, the following sections focus on describing the semantic settling dynamics that would emerge if optimal homonymous and polysemous words that are associated with equally frequent distinct interpretations and contexts were being processed. Insofar as this is not the case, the activation trajectories for the ambiguous words are expected to regress back towards those of a unambiguous words. This may be particularly relevant in the case of homonyms, which previous studies have found to be associated with interpretations that cover a range of frequencies (e.g., Twilley, Dixon, Taylor, & Clark, 1994).

Early context-free processing. The early dynamics in semantics are predicted to be due to the interaction between the inherent processing characteristics of a neurobiologically-inspired computational system and representational differences between unambiguous, homonymous, and polysemous words. For present purposes, the key characteristics of the computational system are that very early processing is dominated by excitatory/co-operative dynamics, whereas later processing is dominated by inhibitory/competitive dynamics. Several factors are assumed to contribute to these dynamics, including: the relatively small proportion of inhibitory versus excitatory neurons in the brain (Armstrong, LeBoutillier, & Petit, 2012; Connor et al., 2006; O'Reilly & Munakata, 2000), which lead to relatively weak and non-specific inhibition that serves to regulate the overall amount of activation in the system (for an explicit demonstration of this principle, see Laszlo & Plaut, 2011; for supporting behavioral evidence and related theories, see Becker, 1980; Khanna & Boland, 2010; Neely, 1991), the presence of exclusively excitatory projections between brain regions, which leads to a transient dominance of excitation when a brain region receives new information, and potentially slower and/or weaker inherent activation dynamics in some inhibitory neurons (Laszlo & Plaut, 2011).

These processing dynamics interact with the representations of ambiguous and unambiguous

words to generate a range of different effects, which are described relative to a baseline consisting of unambiguous words. Setting context aside, the processing of unambiguous words would consist of activating a single set of features that correspond to a distributed representation of the word's meaning (Hinton & Shallice, 1991). Given that these features form a single coherent meaning, all of these features could co-operate to activate one another; for reference later on, this will be refered to as a 'moderate' amount of co-operation. Together, this co-operation will lead to the formation of an attractor basin (Plaut & Shallice, 1993) that gradually leads to the activation of a correct semantic representation. Given the consistency of the intra-semantic mappings relative to the assumed arbitrary nature of the mappings between the surface form of a word (e.g., orthography) and the word's meaning, the contribution of the semantic dynamics to the overall settling dynamics is assumed to dominate any orthographic-semantic dynamics (Plaut & Shallice, 1993), and so the assumed attractor dynamics are expected to result primarily from semantic processing only. The shape of the attractor basin, in conjunction with the nonlinear input-output dynamics of standard connectionist units, is assumed to generate the semantic activation for unambiguous words that is presented in the left portion of Figure 3.1, prior the 'blend state' label, which marks the transition between relatively context-free to contextually-sensitive comprehension. Note that the onset of processing depicted in the figure corresponds to when semantic processing begins and does not reflect assumed early perceptual and orthographic/phonlogical surface form processing of a lexical stimulus. Additionally, the exact rates at which the appropriate and inappropriate interpretations of the ambiguous words are activated should not be taken as specific quantitative predictions and are instead intended to illustrate the broad qualitative predictions of the account.

Relative to the unambiguous words, polysemes are assumed to be associated with two distinct but related senses that have considerable overlap in the semantic features associated with each interpretation. This assumption is consistent with the bulk of current ambiguity research (e.g., Klepousniotou et al., 2008; Rodd et al., 2002; Williams, 1992; but for a range of alterna-

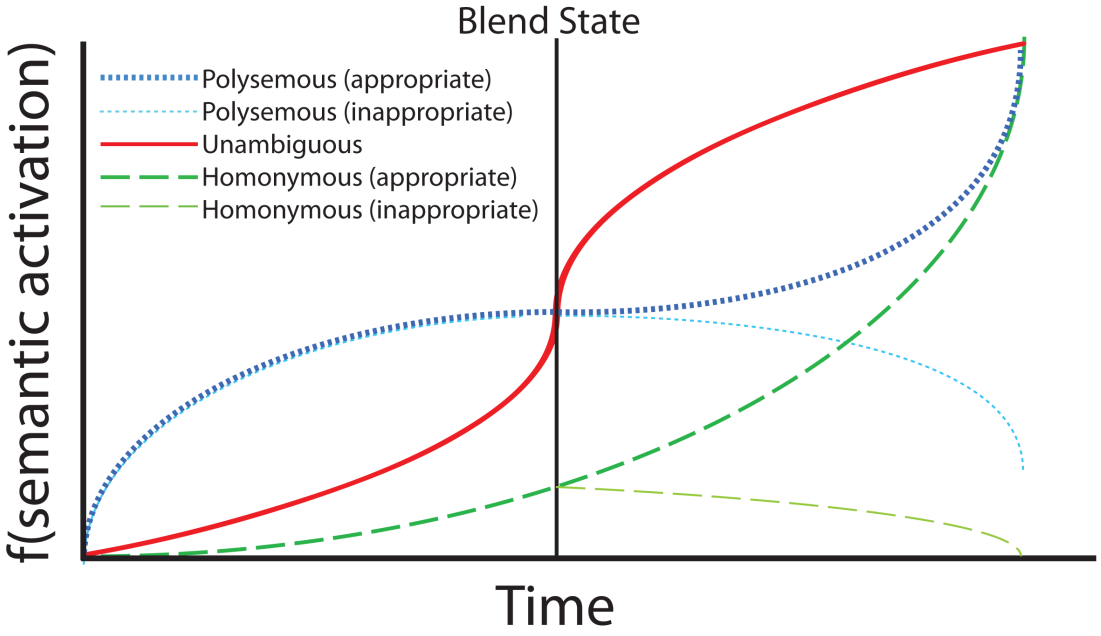


Figure 3.1: Semantic activation trajectories predicted by the settling dynamics account. The plot shows a measure of semantic activity, f(semantic activation), as a function of settling time in semantics for the contextually appropriate and inappropriate interpretations of polysemous, unambiguous, and homonymous words. The blend-state label serves to delineate the early "context-free" settling dynamics from the later "context-sensitive" settling dynamics. The time point beginning at the y-ordinate corresponds to when semantic processing begins and does not reflect assumed early perceptual and orthographic/phonological processing.

tives, see Klein & Murphy, 2001; Lehrer, 1990; Nunberg, 1979). As a result of these partially overlapping representations, polysemes benefit from a high degree of co-operation across their related senses early when processing is dominated by excitatory dynamics. This is because a larger set of features can co-operate to activate the core semantic features that are shared across interpretations. For illustration purposes, these polysemes are idealized as posessing an optimal amount of featural overlap so that they will show a maximal amount of initial facilitation relative to unambiguous controls (insofar as this featural overlap decreases and renders the polysemes more homonymous, or increases, and renders the polysemes more unambiguous, the polyseme trajectories are expected to regress towards the homonymous or unambiguous controls). As processing continues, this early advantage will dissipate as more incompatible features are activated and competition between the representations increases. This would eventually lead to a relatively homeostatic state in which the core features of the polysemes have been relatively strongly activated and the inconsistent features have been partially activated to their statistically optimal values given the absence of context. At this point, assumed to occur near the blend state, the polysemes and the unambiguous controls would not differ from one another in terms of some function of their total semantic activation.

In contrast to the polysemes, homonyms are assumed to be associated with inconsistent and non-overlapping semantic representations corresponding to each of their meanings. Consequently, homonyms would not receive any co-operative advantage for processing early on beyond that present within each of their individual interpretations. Homonyms would therefore elicit the same semantic activation as unambiguous controls, albeit perhaps distributed somewhat differently (e.g., twice as many features activated half as strongly). As features associated with the incompatible meanings are increasingly activated and the influence of inhibitory processes increases, competition amongst conflicting meanings would increase to a very high level and a homonymy disadvantage would emerge later in processing near the blend state.

Late context-sensitive processing. As the influence of context builds up over time and the blend state is approached, the representations of the ambiguous words will gradually be constrained on the basis of their consistency with this context. Upon reaching the blend state, the activation of the semantic representation of an unambiguous word may be slightly facilitated because the unambiguous word is generally assumed to be consistent with the context in which it occurs. However, this facilitation will likely not be very large in magnitude because sufficient information to activate the semantic representation is available from the word form itself and these words do not need to be constrained by a potentially slow and weak contextual influence.

The effects of context are, however, more substantial in the case of ambiguous words. Prior to the blend state, both interpretations of an ambiguous word should have essentially the same level of activation (assuming the interpretations are equally frequent; otherwise, a regression towards the semantic activation trajectory for the unambiguous words is predicted). Following the blend state, the processing of all ambiguous words is slowed as their interpretations compete with one another, with greater competition resulting for homonyms because of the greater degree of inconsistency across their different interpretations. The contextually inappropriate interpretation of a polyseme is not de-activated to the same extent as the contextually inappropriate interpretation of a homonym because some of the features of the contextually inappropriate interpretation of the polyseme are shared by the contextually appropriate and fully activated interpretation.

Depending on the amount of bias towards a particular interpretation of an ambiguous word, a substantial influence of context may occur earlier or later in time. Extremely biasing context (e.g., Tabossi, 1988) may in fact lead to a virtually immediate influence of context from the onset of semantic processing, although such strong biases are viewed as exceptional. The presentation of contexts that are unrelated to an ambiguous word may also lead the system to move towards the interpretation that is most similar to the unrelated context (i.e., more proximal in semantic space), even if this similarity is very low (Rodd et al., 2004), although the semantic features of the consistent interpretation in this case may not be as fully activated as in a strongly biasing

context.

3.2 The Settling Dynamics Account and the Existing Literature

On the basis of a relatively small set of core principles, the settling dynamics account is able, in principle, to generate a wide range of different effects depending on the time point at which the semantic activation in the system influences processing. The following section examines how key effects identified in the literature review could be captured by this account. To be clear from the outset, the goal of this section is not to claim that the full details of each of these effects are perfectly captured by the settling dynamics account and no other representations or processes may be relevant to understanding these effects. Rather, this section serves to evaluate whether, to a first approximation, these effects are at least consistent with the settling dynamics account. Insofar as this is the case, this will be indicative of whether the settling dynamics account could serve as a central pillar for understanding the bulk of the reported effects. In cases in which a detailed account of a particular phenomenon is required, additional representations and processes (e.g., orthography, response selection) may ultimately be relevant, albeit at the expense of parsimony, as discussed in later chapters.

Studies that do not involve context. Experiments that do not involve context are assumed to reflect semantic settling dynamics that occur up to the blend state (Masson & Borowsky, 1995; Piercey & Joordens, 2000; Rodd et al., 2002). In the case of visual lexical decision, responses are typically quite rapid relative to other tasks and relatively coarse and imprecise semantic information would be needed to support a lexical decision. Consequently, visual lexical decisions would be expected to be based on the early portion of the semantic activation trajectories depicted in Figure 3.1 and should be associated with a polysemy advantage, exactly as has been found in

many lexical decision studies (Beretta et al., 2005; Klepousniotou & Baum, 2007; Piercey & Joordens, 2000; Rodd et al., 2002). Auditory lexical decisions are assumed to be based on a similar set of processes but are associated with considerably longer latencies. Assuming that these longer latencies allow for some additional semantic processing to take place relative to visual processing (see, e.g., McClelland & Elman, 1986 vs. Borowsky & Besner, 1993, 2006), the settling dynamics account also captures the homonymy disadvantage and polysemy advantage that have been reported when using this task (Klepousniotou & Baum, 2007; Mirman et al., 2010; Rodd et al., 2002). The regression towards the unambiguous word baseline for words that do not have an optimal amount of semantic overlap or that are not associated with balanced meaning frequencies is also consistent with the non-significant effects of metaphoric polysemes and of unbalanced homonyms (Klepousniotou & Baum, 2007), as well as the effects of noun-verb homonymy (Mirman et al., 2010). The semantic categorization results for tasks that employed broad categories (e.g., living thing) were associated with very slow responses relative to visual lexical decision and were associated with a homonymy disadvantage (Hino et al., 2006). These results are consistent with the later portions of the settling dynamics as the blend state is approached. The semantic categorization tasks that employed narrow categories showed extremely fast responses and may have been based on very early semantic access near the y-ordinate before any of the trajectories had separated to a substantial degree; a similar explanation could also apply to the data from lexical decision tasks that employed relatively easy nonwords and that would consequently not require substantial semantic information (if any) to make a lexical decision (Azuma & Van Orden, 1997; Rodd et al., 2002). The latency increase and subsequent detection of a homonymy disadvantage when the slower go/no go task was employed instead of a standard yes/no semantic categorization task also falls out naturally from the settling dynamics account as the magnitude of the homonymy disadvantage is expected to increase as more time—and thus, more settling in semantics—takes place (Siakaluk et al., 2007). All of the existing neuroimaging data is also consistent with these semantic ambiguity effects emerging in the an-

terior temporal lobes, which are thought of as the central semantic hub in the brain (McClelland & Rogers, 2003), and more tentatively, for the competitive dynamics to be caused at least in part by later activation in the the frontal lobes (in particular, the left inferior frontal gyrus, Hargreaves et al., 2011; Rodd et al., 2010).

Nevertheless, the data from the Hino et al. (2010) lexical decision tasks, in which the characteristics of the nonword foils were varied to cause increases in overall latencies that do not correspond to the predicted effects of the settling dynamics account, represent one study that does not fit well with the current version of the account. As noted in Chapter 2, however, the reliability and generalizability of these findings are questionable for several reasons (e.g., very small item sets, homonymous items that are argually more appropriately classified as hybrid items). The results of the second lexical decision experiment reported in Chapter 5 suggest additional factors that may be relevant to understanding these findings, as well, and so further discussion of these data is deferred to that chapter.

Studies that involve context. Studies that involve the presentation of biasing context require the consideration of the full range of the semantic settling dynamics, both before and after the blend state. The repeated reports of early activation of both interpretations of a homonym, followed by the later activation of only the contextually-appropriate meaning of the homonym (Seidenberg et al., 1982; Swinney, 1979; Tabossi, 1988; Van Petten & Kutas, 1987), are consistent with the semantic activations expected before the blend state and near the end of the plotted activation trajectories. The neuroimaging data also support a semantic locus for these effects (Van Petten & Kutas, 1987). The partial activation of the contextually inconsistent interpretation of a polyseme at the end of processing, due to its shared features with the consistent interpretation, would also explain the subtly different results observed with polysemes (Williams, 1992). The eyetracking data that showed a homonymy disadvantage in the absence of biasing context are consistent with the activation levels predicted at the blend state (which may be as far as processing proceeds before initiating an eye-movement to the next word given the lack of contextual

constraint to guide additional processing), and the disadvantage for both homonyms and polysemes in the presence of a biasing context would be consistent with the activation trajectories observed between the blend state and the final activation levels (Frazier & Rayner, 1990). A similar explanation would also capture the results reported in relatedness judgment experiments (e.g., Piercey & Joordens, 2000).

In its current form, the account does not speak to the developmental trajectories associated with semantic ambiguity (e.g., Booth et al., 2006) or of how these are affected by brain damage (e.g., Taler et al., 2009). However, the past history of the connectionist framework at addressing these types of data suggests that these effects may be within the scope of the theory (e.g., McClelland & Rogers, 2003) and may be emergent characteristics of a system that simulates the details of learning and brain damage.

Summary. Taken together, the settling dynamics account predicts a large portion of the empirical data that have been reported in the complex and often apparently contradictory semantic ambiguity literature. Admittedly, it does not explain every single pattern of results in full detail. Nevertheless, given the lack of a general account of how semantic ambiguity is resolved and the parsimonious explanation of these effects as emerging from simple, independently-supported, domain-general processing and representation principles, this account appears to have much to offer to the field and warrants further investigation.

3.3 Challenges for the Settling Dynamics Account

The generality and parsimony of the settling dynamics account, as well as the account's basic compatibility with a wide range of empirical phenomena, provide considerable initial support for this account as a general theory of ambiguity effects. Nevertheless, as it stands, the account suffers from a number of issues. First and foremost, the underspecified verbal articulation of the settling dynamics account, as it currently exists, raises concerns as to whether the expected

range of ambiguity effects would actually emerge in a system that explicitly implements its assumptions. Related to this point, Hino et al. (2006) specifically argued against an explanation of several task differences based on differences in semantic coding within a parallel distributed (PDP, or alternatively, connectionist) processing system. They assumed that all tasks in which semantic representations of an isolated word are accessed involve the activation of the same semantic code. Consequently, they suggested that the observed task differences "are likely not due to the semantic-coding process as that process is conceptualized within parallel distributed processing models" (p. 266); rather, these task differences must be the result of how the decision-making component of different tasks taps into the semantic code. This is a bold and important claim because connectionist models typically have been shown to provide good accounts of a wide variety of effects related to word processing (e.g., McClelland & Rumelhart, 1981; Plaut et al., 1996) and have shown some basic promise of accounting for multiple semantic ambiguity effects as a result of the nonlinear processing dynamics that can emerge in these types of systems (Kawamoto, 1993; Piercey & Joordens, 2000). The failure of connectionist models to account for basic effects related to semantic ambiguity resolution would therefore be an important constraint on the breadth of effects that can be explained by connectionist modeling that focuses on the development of 'triangle' models of orthographic, phonological, and semantic processing (e.g., Harm & Seidenberg, 1999; Plaut, 1997; Plaut et al., 1996). In particular, such a failure would suggest that considerably more emphasis needs to be placed on understanding the interactions that occur between the lexical-semantic system and the response system.

For these concerns to be alleviated, it is necessary to substantiate the claims of the settling dynamics account via a connectionist simulation that instantiates its core representation and processing assumptions. This will serve to demonstrate that a connectionist system can actually generate the predicted semantic activation trajectories. It will also serve to strengthen the account by generating specific predictions that can be used to guide additional empirical work that evaluates and contrasts the settling dynamics account with other existing theories. Given the

scope and complexity of the account, it is also logical to begin this work by focusing on a subset of the relevant phenomena using methods that can readily be extended in later work to make contact with additional phenomena. From the existing literature, it is a well established and accepted claim that contextual integration is a relatively slow process that leads to different patterns of effects before and after context influences processing (e.g., Seidenberg et al., 1982; Simpson, 1981, 1994; Swinney, 1979; Tabossi, 1988). Additionally, Hino et al.'s (2006) criticism of PDP accounts, and the data they used to support that criticism, are primarily related to processing prior to the onset of a strong biasing context. A reasonable target for the initial simulation work is therefore to explore and understand how the processing dynamics in an explicit system unfold prior to the blend state. The next chapter reports the development of connectionist simulations that are primarily related to this aim.

Another set of issues, related to the first, is the lack of strong *a priori* support for the settling dynamics account and of evidence that specifically supports the predictions of this account and refutes those of other accounts (e.g., accounts based on the response selection system, Hino et al., 2006); thus far, the account has only been shown to be consistent with a set of existing empirical phenomena. Ideally, however, a theory should be suitable not only for understanding and explaining existing data but also for generating predictions that guide future research and advance knowledge by offering insights into how theories can be refined and contrasted to one another (McClelland, 2009; Stanovich, 2001). A central and as-yet unverified prediction of the account is that the total amount of semantic processing that has been completed will be associated with different semantic ambiguity effects. Varying processing time, holding all else constant, should therefore produce a range of different ambiguity effects and provide a strong basis for evaluating and comparing the predictions of the settling dynamics account to theories that posit qualitative, task-specific, underspecified, and ad hoc explanations of particular semantic ambiguity effects (e.g., Hino et al., 2006). Chapter 5 reports the results of several behavioral investigations that examine these issues following the refinement of the predictions of the settling dynamics account

that fell out of the simulation work.

Chapter 4

Substantiating the Settling Dynamics Account: Connectionist Simulations

This chapter reports two connectionist simulations that were aimed at demonstrating how the fundamental assumptions of the settling dynamics account can produce a portion of the theorized semantic settling dynamics outlined in the previous chapter. In particular, this work focuses on the more contentious early portion of the settling dynamics that occur in the absence of strong biasing context (Hino et al., 2006). The first simulation was conducted using a standard set of connectionist assumptions and produced a reasonable—although not perfect—approximation of the theorized semantic settling dynamics. The second simulation employed a more neurobiologically plausible set of connectivity assumptions and the simplest possible representation assumptions that were needed to investigate the expected settling dynamics. This simulation produced an improved set of semantic settling dynamics and allowed for cleaner insight into how the underlying mechanics of the simulation lead to the observed dynamics. Following the presentation of these simulations, the general strengths and weaknesses of this work are discussed and directions for future simulation work are identified.

Note. A portion of the work reported in this chapter was also reported in Armstrong and Plaut (2008).

The key claims that were explored in this simulation work were that the nonlinear dynamics of parallel distributed processing systems are such that different trends can manifest themselves at different time points during processing (Kawamoto, 1993), and that these dynamics would interact with representational differences amongst polysemous, homonymous, and unambiguous words to produce different patterns of effects throughout the course of processing. Thus, the apparent task differences discussed in Chapter 2 may result from the different degrees of semantic precision required to complete each task being associated with different semantic activation levels for the different word classes. For example, very coarse semantic information may be sufficient to decide that a letter string is a word and be influenced by settling dynamics that show a polysemy avantage, whereas semantic categorization may require deriving a sufficiently precise semantic representation to verify category membership and be influenced by settling dynamics that show a homonymy disadvantage.

4.1 Settling Dynamics Simulation 1

The first simulation focused on instantiating the central premises of the settling dynamics account in a simple model of word comprehension using the standard connectionist formalism. In this model, the orthographic representations of words were associated with one or more semantic interpretations that could either not overlap or overlap substantially. The model was then trained to comprehend these words by activating the semantic features associated with the interpretation that was consistent with a context representation. The context representation was presented following the presentation of the orthographic representation of the word to simulate how context generally does not exert an early influence on orthographic-to-semantic mapping, even if that context was established by previously-presented words.

Methods

Network architecture. The model architecture is presented in Figure 4.1. The network was composed of 25 orthographic input units, 75 context input units, 150 hidden units, and 100 semantic output units. The hidden and semantic units integrated their net input over time; their outputs were a sigmoidal function of this net input. The number of hidden units was selected to be as small as possible while still being able to train the network to the training criterion. This was expected to maximize the competition among meanings and senses (see also Joordens & Besner, 1994). Both the orthographic and context units were connected to the hidden units, which in turn were all connected to the semantic units. Additionally, the semantic units were connected back to the hidden layer. Each unit also received a bias connection. For all but these biases and the connections between the context and hidden units, the connection weights were randomly initialized prior to training by sampling from a flat distribution with a mean of 0.0 and a range of 0.3. Given that the training patterns were relatively sparse, the biases were initialized by sampling from a flat distribution with a mean of -3.0 and a range of 0.3 so as to reduce the overall activation in the network at the onset of training. To de-emphasize the importance of context in driving the formation of the initial semantic representations, the context-to-hidden connections were initialized with a mean of 0.0 and a range of 0.05. These connections therefore played a reduced role in driving the activation of the semantic units at the start of training. Consequently, greater performance improvements should be expected from modifying the larger weights from orthography, per the backpropagation rule (Rumelhart et al., 1995).

Training patterns. The training patterns were divided into three groups consisting of 128 unambiguous words, 64 homonymous words, and 64 polysemous words. Each training pattern consisted of an orthographic and context input and a target semantic output. Artificial patterns were generated to approximate the relationship among written words and their meanings. Specifically, all of the representations of orthography, context, and semantics were generated by prob-

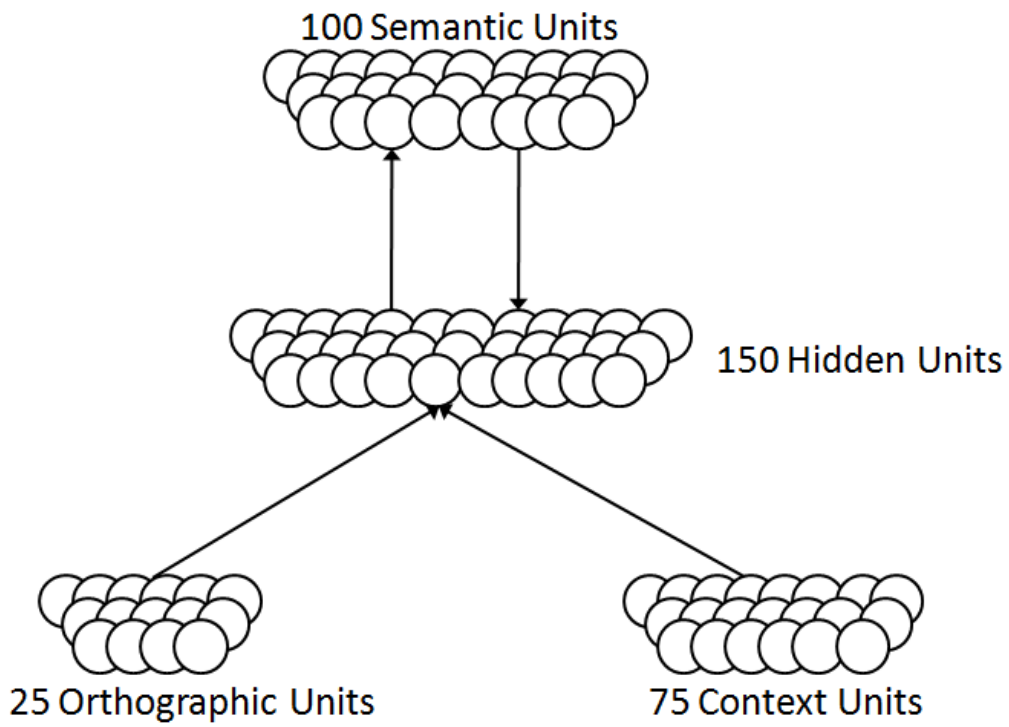


Figure 4.1: Architecture of the connectionist network from the first settling dynamics simulation.

abilistically activating 0.15 of all of the units in the relevant pool of units, with the constraints that at least three units must be active in all patterns, and that all patterns must differ from one another by at least three units. Unambiguous words consisted of a single pairing of a randomly selected orthographic pattern, context pattern, and semantic pattern. The frequency with which this pattern was presented to the network was scaled by a factor of 2.0 so that the orthographic representations of unambiguous words would be presented equally as often as the orthographic patterns of ambiguous words, as in the behavioral experiments (e.g., Rodd et al., 2002; Hino et al., 2006). Homonymous words were represented as two separate input patterns that shared the same orthographic pattern, but were associated with a different randomly selected context and semantic pattern. Polysemous words were represented in a similar manner, except that the semantic patterns for polysemous words were both originally derived from the same prototypical semantic pattern that was permuted so that exemplars of this prototype shared 60% of their

features with one another. The patterns were structured so that the orthographic patterns would appear in isolation for 10 unit updates, prior to the simultaneous presentation of the orthographic and context patterns. The context inputs were soft-clamped to the context units so that their activation would rise gradually and thus integrate smoothly with the state of the network.

Training. The model was trained using recurrent back-propagation through time and a variant of momentum descent in which the length of the pre-momentum weight step vector could not exceed 1.0 (Rohde, 2004). A learning rate of 0.01 and momentum of 0.85 were employed to train the network. Units were considered to be correctly activated once they were within 0.3 of their target activation. Error for units that should be off was scaled by a factor of 15.0, so as to encourage the network to activate only correct units. All of the training patterns were presented to the network in permuted batches. On each trial, error was calculated for the last 5 unit updates. Between each training pattern, the activation in the hidden and semantic units was reset to zero. Training continued until all units in all patterns were on the correct size of 0.5. Training took approximately 6000 sweeps through the training corpus.

Results

The average number of semantic units with activations above 0.7 for the homonymous, polysemous, and unambiguous words at each unit update are depicted in Figure 4.2. Note that these trajectories do not reflect the pre-semantic perceptual processing which is not instantiated in the model; the initial time-step reflects the onset of semantic processing only.

Discussion

The observed activation trajectories map reasonably well onto the theorized semantic settling dynamics described in the previous chapter, and by proxy, to existing behavioral data. Tasks that require little semantic precision (e.g., lexical decision; Figure 4.2: slice A) are predicted to show a polysemy advantage, whereas tasks that require high amounts of semantic precision (e.g.,

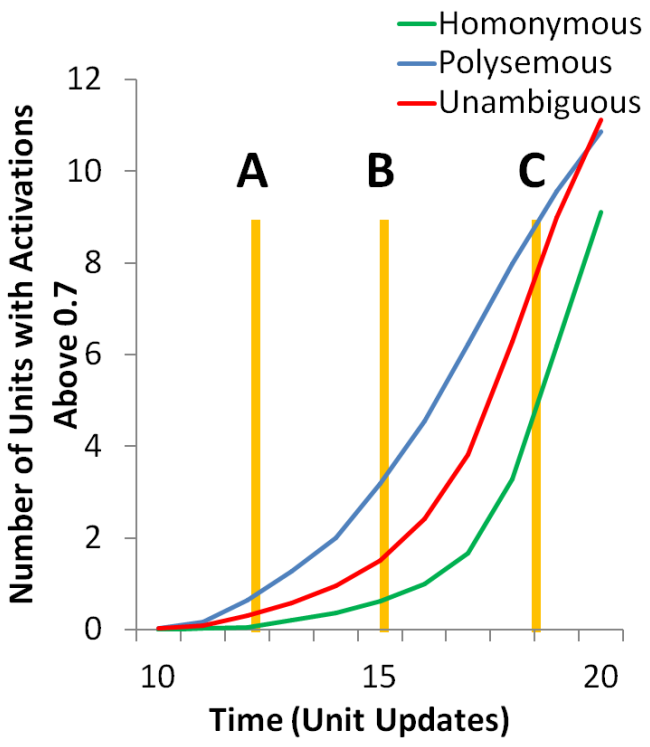


Figure 4.2: The average number of semantic units active above 0.7 for polysemous, unambiguous, and homonymous words in the first settling dynamics simulation. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only. No semantic units were active above 0.7 before unit update 10. Slice A: polysemous words settle more quickly than unambiguous words, which in turn settle fractionally more quickly than homonymous words. Slice B: Theoretical cross-over point at which the trajectories for polysemous words and homonymous words are both significantly different form unambiguous words. Slice C: Polysemous words are fractionally faster than unambiguous words, and both are substantially faster than homonymous words. Vertical as opposed to horizontal slices are used because the claim is only that semantics contributes to driving performance in these tasks, not that the current model is a comprehensive account of the dynamics underlying the different behavioral tasks as a whole.

semantic categorizations involving broad categories; Figure 4.2: slice C) are predicted to show a homonymy disadvantage. Furthermore, tasks that require moderate semantic precision are predicted to show both a homonymy disadvantage and a polysemy advantage (Figure 4.2: slice B), and tasks that require either extremely high or extremely low amounts of semantic precision should show no differences between the word conditions, potentially explaining some observations of null effects of ambiguity (e.g., Azuma & Van Orden, 1997; Klein & Murphy, 2001). At a general level, the model's behavior thus supports the notion that the processing dynamics in the semantic representations of the word comprehension system could be the primary cause of the disparate empirical findings reported in the literature. Additionally, this work produced a novel prediction that both a polysemy advantage and homonymy disadvantage should be observed for tasks that tap an intermediate time-point in the settling dynamics prior to the blend state.

Nevertheless, there are several issues with this simulation. Some of these issues are shared with the second simulation and their discussion is defered until after the presentation of that simulation. Of the remaining issues, the foremost is that the model failed to fit certain aspects of the empirical data. Although at a quantitative level the polysemes showed a greater processing advantage than the homonyms early in processing and the converse was true later in processing, this does not reflect the patterns of significance observed in the behavioral data. In those data, the early polysemy advantage is typically detected in the absence of any homonymy disadvantage whatsoever—not a homonymy disadvantage that is lesser in magnitude than the polysemy advantage.

A second issue with this simulation is that the number and complexity of the distributed representations—although simple relative to the actual representations they are intended to stand for—precludes a straightforward assessment of the underlying principles that give rise to these effects. To conduct such an assessment, it would be useful to further simplify the simulation in some respects, given that these principles have been shown to operate reasonably accurately in this somewhat complex simulation, to more clearly explore the detailed mechanics that are

responsible for these settling dynamics. The second simulation aimed to address both of these issues.

4.2 Settling Dynamics Simulation 2

The second simulation was similar to the first with two important differences. First, the orthographic, context, semantic representations that the simulation was trained on were simplified to the maximum extent possible to facilitate the assessment of the mechanics of the system and how they produced the observed semantic settling dynamics. Second, the weights that connected the units in the simulation were constrained on the basis of the structure of the connections and distribution of excitatory and inhibitory neurons in cortex (Armstrong, 2007; Connor et al., 2006; Kandel, Schwartz, & Jessell, 2000; O'Reilly & Munakata, 2000). These constraints were broken down as follows.

The most important characteristic of the neurobiological constraints was that the units in this simulation needed to reflect two distinct populations of excitatory and inhibitory neurons. Consequently, an individual unit could only send excitation (positive weights) or inhibition (negative weights)—not both. The proportion of excitatory versus inhibitory neurons is also highly unequal in cortex such that only approximately $10-20\%$ of neurons are inhibitory (Armstrong, LeBoutillier, & Petit, 2012; Connor et al., 2006; O'Reilly & Munakata, 2000). This may serve to bias the types of computations to which these different populations of neurons contribute, such that excitatory units contribute to the rich and detailed computations that activate particular features whereas the inhibitory units only contribute to regulating the overall activity in a given system. Moreover, to a first approximation the connections between different brain regions—for instance, the occipital and posterior temporal lobe regions subserving orthographic representation and processing and the anterior temporal lobe regions that subserve semantic representation and processing—are excitatory only whereas both excitatory and inhibitory connections are present within a brain region (Kandel et al., 2000). As a result, one hypothesized characteristic

of this connectivity structure is that early processing will be dominated by the initial excitation that enters the system and spreads amongst consistent semantic features, whereas the later dynamics will be more strongly influenced by inhibition once the inhibitory units are activated to prevent the over-activation of the brain region. Furthermore, the density of the axonal projections between brain regions is relatively sparse whereas the density of projections within a brain region is generally quite dense (Laughlin & Sejnowski, 2003). For present purposes, these density differences are approximated as weight strength differences, such that the average magnitude of the weights within a brain region are biased to be stronger than those between brain regions. This allows for the development of a smaller simulation that approximates the average relative inputs from within and between brain regions. A relevant characteristic of this type of connectivity structure is that most of the constraints on semantic feature activation may be provided by the strong intra-semantic connections whereas the connections into semantics from an orthographic or context representation will impose only weak constraints on settling in semantics. These constraints should also exacerbate the relatively weak constraints imposed by the arbitrary similarity structure inherent to orthographic-to-semantic mappings and reinforce the potentially strong constraints imposed by intra-semantic mappings (Plaut & Shallice, 1993). This should further simplify the interpretation of the semantic settling dynamics as resulting primarily from semantic processing, despite the fact that the simulation framework is inherently interactive and activation flows continuously across all of the weights.

Methods

Network architecture. The model architecture is presented in Figure 4.3. The network was composed of four orthographic input units, each of which was used to represent an individual word (i.e., localist orthographic representations were employed), two context units, each representing one of two separate contexts, and 22 units representing semantics, which integrated their inputs over time and produced outputs that were a sigmoidal function of their net inputs. Of these

22 units, 21 were excitatory units and were used to code the distributed semantic representations of the words presented during training (discussed next), and 1 was an inhibitory unit that regulated activation in semantics and stood for an inhibitory interneuron. Only a single inhibitory unit was employed in the present simulation, despite the fact that $2-4$ units would better approximate the actual distribution of excitatory and inhibitory neurons in the brain, because the fundamental assumption of this division is that the information-content that can be encoded via the inhibitory units' weights should be minimal. In the present simulation, however, even two inhibitory units could, together, represent critical information content (e.g., by each specializing in providing constraints associated with one of the two different contexts, as was observed in a pilot simulation). Consequently, employing only a single inhibitory unit, although not in agreement with distribution of excitation and inhibition itself, better approximates the assumed contributions of these populations of neurons to the computations that are carried out within a brain region.

Both the orthographic and the context units were connected to the semantic units. The representation of orthography was assumed to be in a different brain region than the semantic representation; a similar assumption was made for the contextual knowledge (which may involve, e.g., a working memory component, Mason & Just, 2007). Consequently, these between brain region connections were restricted to be excitatory only and the weights were constrained to never be less than zero in magnitude. The semantic units were also connected to one another with the exception that these units were not connected to themselves. Self connections were not permitted because self-excitation is assumed to be matched or outweighted by leakage in actual neurons (Usher & McClelland, 2001). Additionally, the relative contribution of a neuron's self-excitatory connection was assumed to be miniscule compared to its input from other neurons contributing to the rich semantic representation of an actual word. Only the inhibitory unit was constrainted to send negative weights to the other semantic units; all of the other units sent positive weights. All of the excitatory weights were initialized to a value of 0.05. The inhibitory weights were

initialized to a value of -0.2. The initial value of the inhibitory weight was set so as to balance, in approximate terms, the amount of excitation that a semantic unit was expected to receive early in training (4 excitatory semantic units expected to be activated per pattern $\times$ 0.05 = 0.2) with an equivalent amount of inhibition (1 inhibitory semantic unit $\times$ $-0.2 = -0.2$). Additionally, all of the units in the network had a bias connection that was set to an initial value of 0. There was no variability in the weight values other than that described above to simplify the interpretation of the weight structure of the network.

To constrain the magnitudes of the weights that were allowed to emerge within versus between brain regions, the between-region connections were subjected to weight decay after each weight adjustment ($\lambda = 0.00025$). The within-region connections (including the connections to and from the inhibitory unit) were not subject to any weight decay. Note that in contrast to the first simulation, the incoming weights from the context representation were not biased to be smaller than those from orthography, and the magnitude of the weights were equally small both between and within brain regions. Consequently, differences in the relative magnitudes of the weights in the trained system should emerge only if this is adaptive during training given the imposed constraints.

Training patterns. The orthographic and context representations employed in this simulation were localist and each unit in those pools of units stood for the representation of one word form or context, as applicable. Additionally, although a distributed semantic representation was associated with each of the words, the semantic features associated with a given word were associated only with that word. This representation scheme was adopted for several reasons. First, it orthogonalized the input/output associations for each word in a way that approximates the lack of similarity structure in orthographic-to-semantic mappings. Second, employing localist representations also avoided the need to introduce variability in the initial weights that is typically employed to overcome the lack of input/output similarity structure between orthography and semantics (and in other simpler similar problems such as XOR). Third, these implementa-

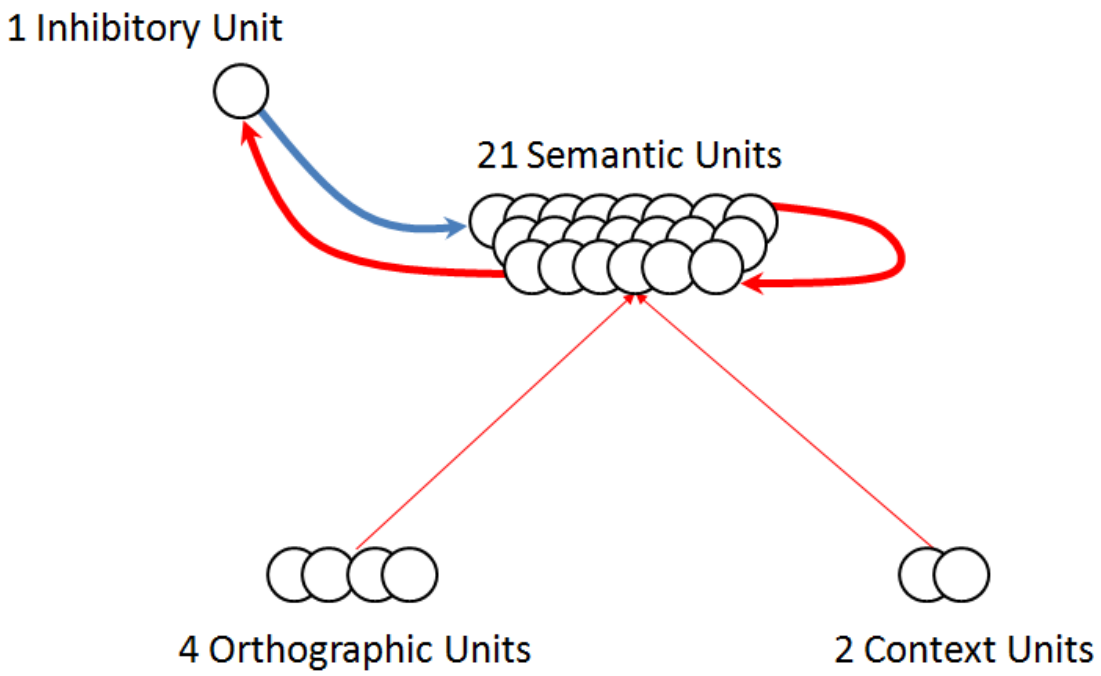


Figure 4.3: Architecture for the network from the second settling dynamics simulation. Red arrows indicate positive (excitatory) connections and blue arrows indicate negative (inhibitory) connections. The thickness of the arrows indicates the constraints on the magnitudes of the weights that were imposed via weight decay; larger arrows indicate that the weights were less constrained and could be adjusted to larger values.

tion choices allowed for a simpler and more direct initial and final weight structure that would be easier to interpret visually and serve to compliment the results of the first simulation that employed richer representations at the expense of ease of understanding of the underlying network mechanics.

The network was trained to represent four different words that were divided into two unambiguous words and two ambiguous words. Each unambiguous word consisted of the pairing of a single orthographic unit with four semantic units. These orthographic/semantic input/output pairs were presented in two different contexts, represented as the activation of one of the two contexts. The ambiguous stimuli consisted of one polyseme and one homonym that were presented in two different contexts. The pattern structure for these representations was the same as for the unambiguous words, with the following exceptions: For the homonyms, four different semantic units were associated with the orthographic input depending on which context was activated. For the polysemes, three semantic units were associated with the orthographic input regardless of the context, but the fourth semantic unit was different for each context. The patterns were structured so that for each pattern the orthographic input would be activated for 20 unit updates, after which both the orthographic and context outputs were activated for an additional 20 updates.

Training. The network was trained using recurrent back-propagation through time and momentum descent. A learning rate of 0.0001 and momentum of 0.9 were employed (momentum was set to 0 for the first update). A relatively small learning rate was selected to maintain a positive gradient linearity across the weights that resulted from each weight update; this value also helped avoid large variations in error that occured with greater probability when larger learning rates were used. Units were considered to be correctly activated when they were within 0.15 of their target activation. Error was computed for the last 5 unit updates. Note that in contrast to the first simulation, error for units that should be off was not scaled differently than for units that should be on. All of the training patterns were presented to the network in permuted batches. Between each training pattern the net input was reset to -1.4 and the net output was reset to 0.2

for all of the units. Training continued until all units in all training patterns were on the correct side of 0.5 for the last 5 unit updates. Training was complete after approximately 55,000 sweeps through the training corpus.

Results

Settling dynamics. The number of semantic units with activations above 0.75 are presented in Figure 4.4. The figure shows an early polysemy advantage (Slice A), an intermediate and short-lasting polysemy advantage and homonymy disadvantage (Slice B), a later homonymy disadvantage (Slice C), and a very late disadvantage for all ambiguous words relative to unambiguous words (Slice D).

To evaluate the generality of the results in a more sensitive and continuous fashion, Figure 4.5 plots the amount of semantic activation that exceeded a value of 0.5 (i.e., the critical value that delineated whether a feature was absent or present for the purposes of training). To a first approximation, this measure may be thought of as expressing the confidence or strength with which a given semantic feature has been activated given that it should be activated, with activations below this level essentially being discarded as subthreshold noise (for similar assumptions associated with the response threshold in a more biologically-plausible model of response selection, see Usher & McClelland, 2001). This figure recapitulates the results from the previous figure.

Underlying mechanics. The weight structure of the network was examined to determine how the network's 'memory' for these words, as encoded in the weights, could interact with the representations of the different words to generate the observed semantic settling dynamics. The weight structure is presented as a 'Hinton diagram' in Figure 4.6. A simpler approximation of these results is presented in Figure 4.7 for ease of reference. The semantic features corresponding to the distributed representation of each interpretation of a word are adjacent to one another. First, note that in the case of each word class, the features that correspond to each interpretation

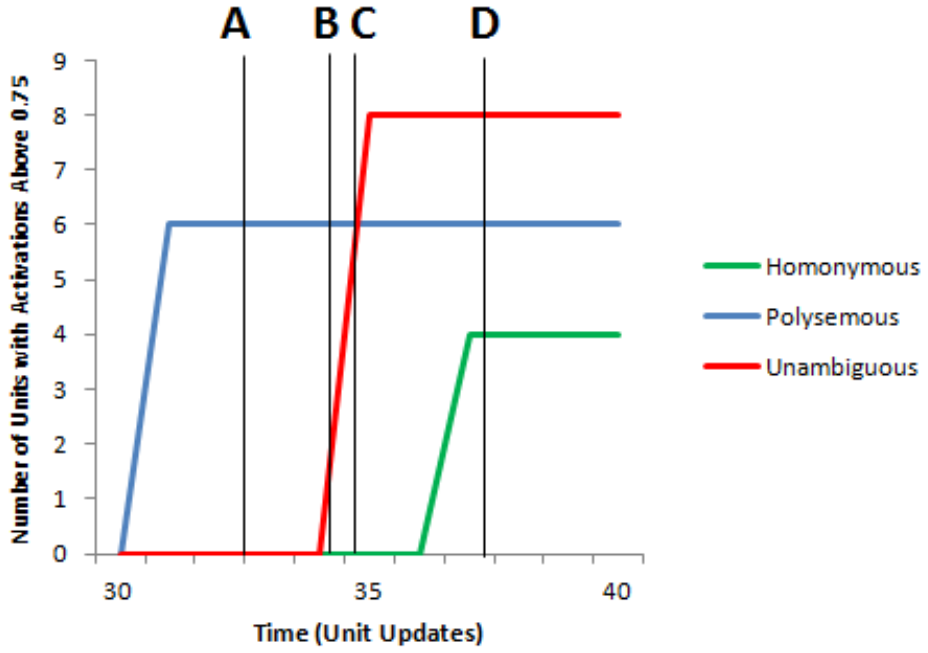


Figure 4.4: The number of semantic units with activations above 0.75 for polysemous, unambiguous, and homonymous words in the second settling dynamics simulation. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only. No semantic units were active above 0.75 before unit update 30. Slice A: polysemous words are settling more quickly than unambiguous words, which in turn are settling fractionally more quickly than homonymous words. Slice B: Theoretical cross-over point at which the trajectories for polysemous words and homonymous words are both significantly different from unambiguous words. Slice C: A reversal of the ambiguity advantage occurs; polysemous words are fractionally faster than unambiguous words, and both are faster than homonymous words. Slice D: A processing advantage for unambiguous words is observed relative to both polysemous and homonymous words.

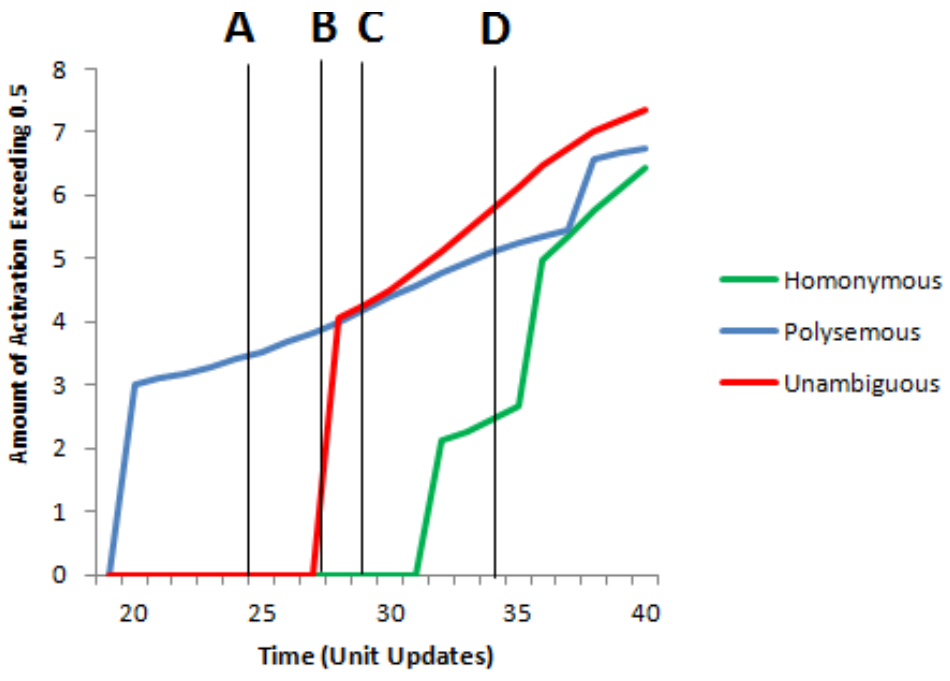


Figure 4.5: The average amount of semantic activation exceeding a threshold of 0.5 for polysemous, unambiguous, and homonymous words in the second settling dynamics simulation. Note that these trajectories do not reflect pre-semantic visual and orthographic processing; the zero time-point reflects the onset of semantic processing only. No semantic units were active above the 0.5 threshold before unit update 19. Slice A: polysemous words are settling more quickly than unambiguous words, which in turn are settling fractionally more quickly than homonymous words. Slice B: Theoretical cross-over point at which the trajectories for polysemous words and homonymous words are both significantly different from unambiguous words. Slice C: Polysemous words are fractionally faster than unambiguous words, and both are faster than homonymous words. Slice D: A processing advantage for unambiguous words is observed relative to both polysemous and homonymous words.

of a word form strong mutually-supporting excitatory sub-networks. In the case of the polysemous words, slightly weaker mutual support originates from the distinguishing features that are unique to each interpretation relative to the core features shared across interpretations, and the inconsistent features do not support one another. Numerically, the weights amongst the core features were higher for the polysemes (weight value = 2.93) than either the homonyms or unambiguous words, which had approximately equal intra-semantic excitatory weights (2.39 and 2.37, respectively). The input from the inhibitory weight was largest for the homonyms (-2.21), followed by the input to the core features of the polysemes (-2.00) and unambiguous words (-1.85); the least inhibition was associated with the distinguishing features of the polyseme (-1.66). This is generally consistent with the different degrees of competition that were expected in semantics for different numbers and relatedness amongst interpretations. The bias connection was weakest for the homonyms (-1.37) than for the unambiguous words (1.49) or polysemes (1.57), although the range of the differences in bias weights was relatively small compared to the range of differences amongst the intra-semantic weights. Taken together, these weight values indicate that the polysemes, and in particular their core features, are associated with greater degrees of excitation than the other word classes and that these dynamics drive the polysemy advantage—particularly during early processing before the inhibitory unit is activated to a substantial degree. Subsequent to the strong activation of the inhibitory unit, the homonyms receive the strongest inhibition, consistent with the homonymy disadvantage.

The incoming weights from orthography and context were generally at least an order of magnitude smaller than the weights within semantics. This is consistent with the objective of developing a network with larger within- versus between-region weights, as inspired by the neuroscience literature, and suggests that these dynamics play a relatively small role in processing. Nevertheless, the role of these representations in driving early semantic activation and in supplying necessary information for the activation of a contextually appropriate interpretation warrants their detailed examination. With respect to the orthographic inputs, the strongest out-

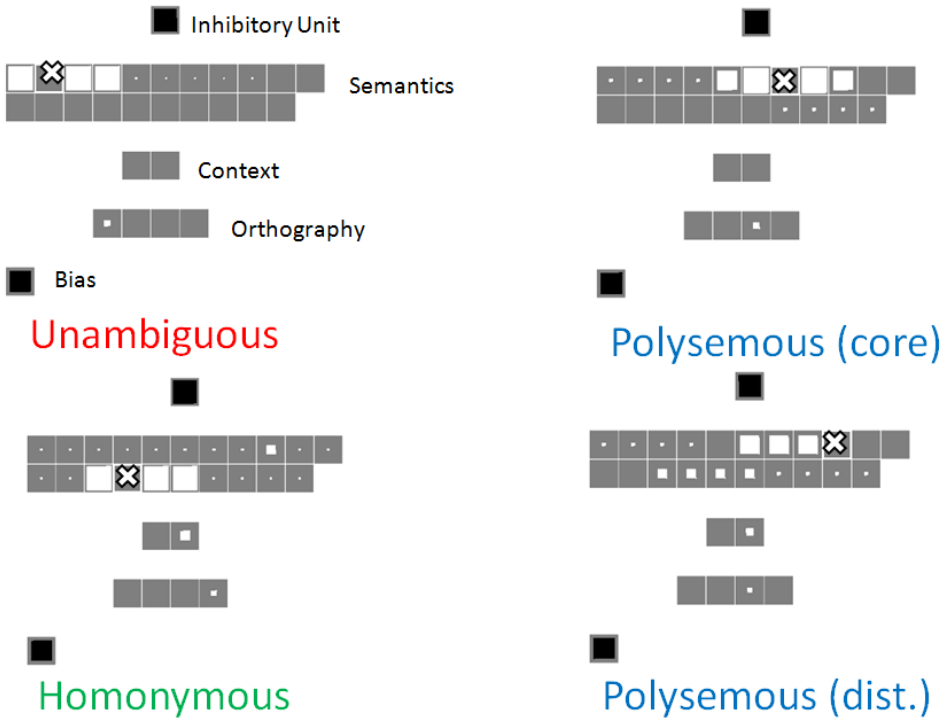


Figure 4.6: Plot of the incoming weights for one interpretation of each type of word that the second simulation was trained on. The 'x' marks a target unit to which all of the other depicted units send incoming connections. Two plots are presented for the polysemous words: one corresponding to the incoming weights for a core feature shared across senses, and one corresponding to the incoming weights for a distinguishing feature associated with only one of the two senses. The color of the square indicates the sign of the weight: Black squares correspond to inhibition and white squares correspond to excitation. The size of the square corresponds to the magnitude of the weight, with larger squares indicating larger magnitudes.

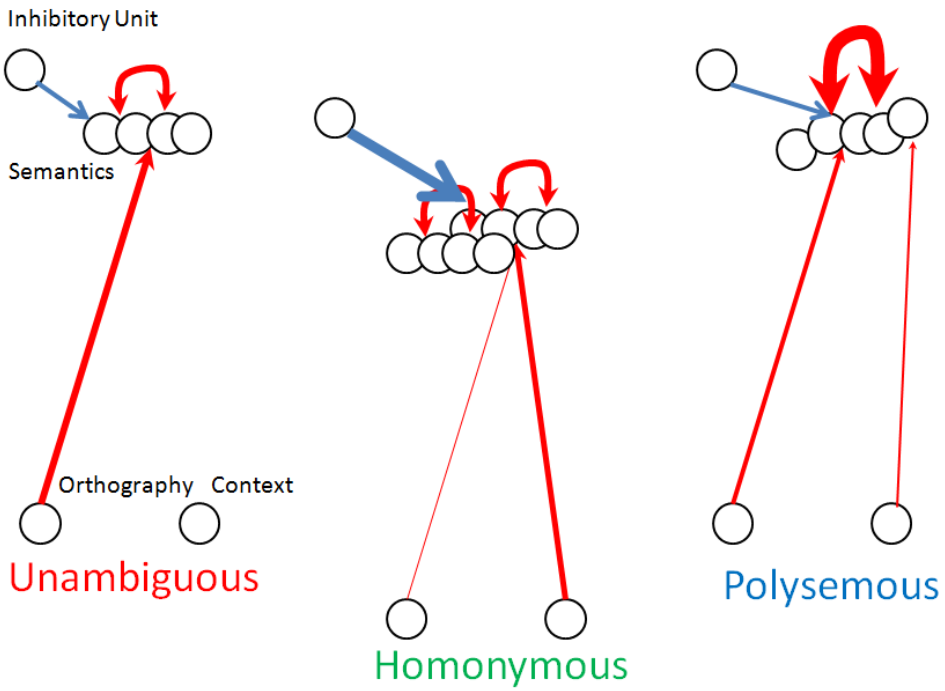


Figure 4.7: Summary illustration of the weight diagram presented in Figure 4.6. The distinguishing features of the polyseme appear as a single group of three units and the two distinguishing features are offset. Red indicates positive weights to the group, blue indicates negative weights to the group, and the thickness of the line reflects the approximate magnitude of these weights relative to other between- or within-region weights; the only exception is the context projection to the distinguishing features of the polyseme, which represents only the inputs to the distinguishing features (the incoming weight to the core features was zero).

going weights were to the core features of the polyseme (0.11), followed closely by the weights to the unambiguous word's semantic features (0.10), by the weights to the semantic features of the homonym (0.08), and finally by the weights to the distinguishing features of the polyseme (0.05). With respect to the context units, the semantic features of the unambiguous words and the core features of the polysemous word were not sensitive to this information, despite its potential usefulness (weights = 0). The distinguishing features of the polysemes had moderate incoming weights from context (0.14); the semantic units associated with the homonyms were also substantially influenced by a relatively strong weight from context (0.25). Taken together then, these results support the theorized mechanics that were expected to give rise to different effects throughout the course of settling in semantics.

Discussion

The results of the second simulation provide further support for the theoretical semantic settling dynamics outlined in the second chapter. In particular, these simulations showed a clear polysemy advantage in the absence of a homonymy disadvantage early in processing, followed by both a homonymy disadvantage and a polysemy advantage (a novel prediction of the simulations), a homonymy disadvantage in the absence of a polysemy advantage, and finally a general ambiguity disadvantage. Furthermore, the constraints employed to generate these predictions and address some of the limitations of the first simulation are quite parsimonious given their origins in independent research into the neurobiology of the neocortex. Admittedly, the time-course of the ambiguity effects predicted by the second simulation may be overly compressed towards the middle section and the relative magnitudes of the different effects may be incorrect in the details. However, these characteristics of the simulation should not be viewed as strong predictions given the small scope of the simulation, which was primarily intended to illustrate the key qualitative characteristics and allow for the examination of the mechanics that support them.

With respect to these mechanics, the second simulation succeeded at providing detailed and

straightforward insight into the weight structure that emerges during training, and how it contributes to generating the observed settling dynamics. In particular, the results are generally in line with the underlying assumptions of the settling dynamics account regarding co-operation and competition within semantics. Furthermore, this simulation points to a further refinement of the account that would involve the small but important relative contributions of orthography and context to shaping the settling dynamics in semantics.

4.3 General Discussion

Taken together, the two simulations provide important support for the underlying assumptions and predictions of the settling dynamics account. Furthermore, these explicit explorations of the core principles of the account generated a novel prediction—the presence of both a homonymy disadvantage and polysemy advantage in tasks that require an intermediate amount of semantic precision. Furthermore, the simulations point to a potentailly non-trivial role for orthographic and context inputs themselves in shaping the settling dynamics in semantics.

Of course, these simulations are not perfect recapitulations of the settling dynamics illustrated in Chapter 3. A number of reasons may explain these discrepancies. Optimistically, some of these differences may correspond to refinements of the verbal articulation of the theory that result from explicitly building a system that implements the theory's core assumptions. However, these discrepancies may also be due to the simplifications adopted in the present work or to details of the underlying neurobiology that are not captured by the present account and that are still active areas of research (e.g., the net input/output function for inhibitory populations, Laszlo & Plaut, 2011). Additional work will be needed to evaluate these possibilities.

Finally, despite the relative success of the simulations it is important to note their limitations and how they may be addressed in future work. One of these limitations is concerned with whether semantic settling dynamics are actually the basis upon which decisions are made. The discussion of this issue is set aside for the general discussion at the end of Part I because of its

overarching relevance to the account and its close relation to the work described in Part II of the dissertation. The second limitation is concerned with how context was treated in the present simulations and is discussed in detail below.

Future simulation work: Improved context representations. A fundamental assumption of the settling dynamics account is that contextual influences are weak during early processing and gradually increase throughout processing. In the simulations reported here, the assumption of weak early contextual influences was implemented explicitly by not presenting a separate contextual input until after the orthographic input had been presented. However, this simplifying assumption is suboptimal because it does not explain how or why the influence of context is weak. The following section evaluates how reasonable this assumption is for present purposes and how the representation of context could be better treated in future work.

A simple means of 'explaining' the weak late effect of context would be to explicitly formulate the settling dynamics theory as a staged theory of word versus discourse comprehension, wherein the computation of a word's interpretation(s) occurs in a context-free fashion prior to deriving a context-sensitive representation. Indeed, such a staged theory of isolated-word and context-sensitive processing was articulated by Swinney (1979). However, such a theory leaves something to be desired because it does not explain how or why such a staged integration would be beneficial—particularly within the context of brain-style computation, which is highly interactive in nature. Rather, such an 'explanation', without further treatment, amounts to no more than a redescription of the data and highlights why failing to discuss this aspect of the simulation methods would introduce a critical, potentially invalid prior (in the standard bayesian sense of the term) to the model (for discussion of this issue, see McClelland et al., 2010). In particular, if the system is interactive and disambiguating context is presented before an ambiguous word (e.g., RIVER BANK), it is important to jusify why the effects of context would nevertheless be weak during early processing.

Several underlying factors may explain this discrepancy. First, to comprehend language the

brain must rely heavily on the orthographic inputs themselves so as to represent and process sentences that are sensible but that violate established context (e.g., 'While fishing, I remembered I needed to go to the BANK once I was done to withdraw some money'). Consequently, it is reasonable to assume a primacy of the stimulus itself, which would initially lead to at least the partial activation of all interpretations regardless of context. Similarly, it may, on average, be beneficial to partially activate all of the interpretations of a word because doing so is beneficial when biasing context has not yet been presented (Mason & Just, 2007; McClelland et al., 1989). Second, by analogy to the visual system, it is possible that the complexity of semantic representation gradually increases throughout successive processing layers so as to gradually transform from a representation of isolated words to a representation of the scenario that the stream of words is intended to evoke (McClelland et al., 1989). This would not constitute a staged theory of processing in the classic sense of Sternberg (1969), but would reflect a specialization of representation that may occur at multiple levels by virtue of the nature of the architecture of neocortex and of the environment (i.e., language that is broken down into individual words that can be combined to denote something more than the sum of each individual word's interpretation). Such a staged theory is also expected to some degree given the role of distal working memory structures (Mason & Just, 2007) and frontal structures (Hargreaves et al., 2011; Rodd et al., 2010) in the disambiguation process. Third, the influence of context, even if it has already been established by previous discourse and has been propagated back to an early represenation to constrain it from the onset of processing, may be weak because it is imprecise. Indeed, if a perfect context could be established for a given word it is reasonable to hypothesize that the word would not be articulated in many cases because it would not contribute any additional information to the sentence. This claim is supported implicitly by the substantial effort needed to identify highly biasing contexts associated with high cloze probabilities for a target word (e.g., Van Petten & Kutas, 1987). A consequence of such an imprecise context would be that although the semantic representations may be slightly biased towards activating the features associated with a given

word, these features would not be activated very strongly and the features of many other words would need to be activated as well. It would not be until substantial word form information had been sent from orthography that the contextual constraint would exert a strong influence on processing to activate the interpretation of one particular word in a particular context.

In all likelihood, all of these different factors contribute to some extent to the weakness of context and serve to justify, in the interim, the present simulation work, which involved a direct manipulation of when contextual information arrived in the system. Ultimately, however, these assumptions warrant further investigation in additional simulations. One particular avenue for these investigations is to implement a deep recurrent network (similar in principle to the networks developed by Hinton & Salakhutdinov, 2006, without such a strong emphasis on data reduction) that does not have its activation states reset between the presentation of words and that does not have an explicit representation of context. This network could then be trained to activate a contextually-appropriate semantic representation of a given word on the basis of the previous words that had been presented, mapping between the input and output through multiple hidden layers. In this case, the contextual information would have to derived internally on the basis of the semantic (i.e., featural overlap) and lexical (i.e., word form association / transitional probabilities) relationships amongst the previous words (McClelland et al., 1989; Plaut, 1995). Higher-level bayesian and statistical modeling investigating how the words in language are roughly grouped to represent different topics of knowledge (Griffiths, Steyvers, & Tenenbaum, 2007), and into the similarity structure of lexical-semantic representations within a category (Johns & Jones, 2010), could also be adapted to provide suitable representations for the simulation. The different hidden layers, as well as the output layer, could then be assessed in the trained network to evaluate whether the influence of context increases as distance from the input increases and distance to the output decreases. This development of such a simulation would, obviously, be quite a substantial undertaking and preliminary work in this regard has been challeged in part by the scale of simulation that is required to do justice to implementing weak

context as an imprecise and non-specific context, but this work has also shown initial promise at shedding additional light on these issues. The assessment of the context-free versus the context-sensitive nature of the representations that may develop in the hidden layers of such networks and how they may contribute to response selection, as well as to the development of an improved set of biologically-plausible processing assumptions, would also assist this work. Part II of the dissertation provides some insights into some of these issues.

Chapter 5

Evaluating the Settling Dynamics Account: Behavioral Investigations

The simulation work established that the key principles of the settling dynamics account operate as hypothesized but in and of itself that work showed only that the settling dynamics account is consistent with past data. The following chapter aims to provide stronger support for the settling dynamics account by testing one of its key predictions: Variations in the degree of semantic precision required to generate a response in a given task will determine the detailed patterns of ambiguity effects that are observed. Evaluating this hypothesis will thus test the utility of this account in guiding future semantic ambiguity research. It will also provide a valuable data set that discriminates between the settling dynamics account and an account of semantic ambiguity effects based on the configuration of the decision system (Hino et al., 2006). To preview the results, a lexical decision study that extended the work of Rodd et al. (2002) by manipulating nonword difficulty initially appeared to have produced the predicted pattern of effects. However, more detailed analyses of these data identified several issues with this study that forced this con-

---

Note. A portion of the work reported in this chapter was also reported in Armstrong and Plaut (2008) and Armstrong and Plaut (2011).

clusion to be rescinded. To address these issues, stimulus optimization software and improved norming methods were developed and used to design an improved item set. A lexical decision task that employed this item set and attemped to vary processing time via both nonword manipulations (orthographically wordlike nonwords vs. very orthographically wordlike nonwords vs. pseudohomophones) and stimulus contrast manipulations (white-on-black vs. grey-on-black) produced effects that were broadly, albeit not perfectly, consistent with the settling dynamics account. Future work that could expand on these studies is discussed in the last section.

Before turning to the experiments themselves, however, it is worth briefly outlining the underlying motivations for running these particular experiments. As noted earlier, the notion that processing time is a critical factor in explaining semantic ambiguity effects is already accepted in principle by many researchers who have investigated the time-course of contextual biases on processing (e.g., Simpson, 1981; Swinney, 1979; Tabossi, 1988). In contrast, the notion of similar time-course effects during early processing prior to a strong contextual bias has been strongly opposed. In particular, Hino et al. (2006; see also Hino et al., 2010) have claimed that an identical semantic code is tapped by a range of tasks and consequently that the different ambiguity effects that have been observed—specifically those associated with lexical decision and semantic categorization—are due to different configurations of the decision system across these tasks. The most controversial aspect of the settling dynamics account is thus associated with the early processing dynamics that it predicts, which contrast quite strongly with the predictions of Hino and colleagues.

A strong empirical evaluation of the settling dynamics account that would also generate predictions that distinguish it from the decision system account described by Hino et al. (2006) is to hold task—and by proxy, the configuration of the decision system—constant and vary only the precision of the semantic code that is associated with a response. Given Hino et al.'s focus on semantic categorization and lexical decision, a demonstration of these effects that involved these two tasks would be the most direct and transparent method of contrasting the empirical and theo-

retical bases of these two accounts. To this end, Armstrong (2007) ran a semantic categorization task similar to that reported by Hino et al. using English as opposed to Japanese (Katakana script) stimuli. These stimuli were selected from the very large set of concepts reported by McRae, Cree, Seidenberg, and McNorgan (2005) that had been the subject of extensive norming and that would consequently allow for relatively tight control over potentially-confounding variables (e.g., semantic richness, as assessed by total number of features, Cree & McRae, 2003; Klepousniotou & Baum, 2007). These items were selected such that the different stimuli conformed to a taxonomic hierarchy that would allow for the development of semantic categorization tasks involving varying degrees of category breadth (e.g., living things vs. nonliving things, animals vs. plants). The aim of the study was to gradually vary the breadth of the category and of overall latencies, and by proxy, the presumed amount of semantic settling involved in this task. Subsequently, these same items were to have been used in a complimentary lexical decision study similar to those reported later in the chapter.

These semantic categorization experiments did not, however, succeed in producing any differential effects of ambiguity once relevant confounds had been controlled for using multiple regression, despite sampling from a population of stimuli that were arguably the most suitable for this task that were available in English.[1] One possible explanation for this lack of ambiguity effects was that a much smaller set of homonyms is available to choose from in English than in Japanese. This is because Japanese Katakana script is employed primarily to represent foreign (and in particular, English) words in Japanese and conflates a number of phonemes that are typically distinct in other languages (e.g., the lack of an R/L distinction causes RIGHT and LIGHT to share a surface form in Katakana and thus create a new homonym). Consequently, it may be simpler, in principle, to conduct the desired manipulation of semantic precision within

[1]Recently, Hargreaves et al. (2011) succeeded in replicating the homonymy disadvantage reported by Hino et al. (2006) in a semantic categorization task that employed English items. However, far fewer items were employed in this task than in the Hino et al. or Armstrong (2007) studies. Consequently, this effect does appear to be present in both languages, although it is somewhat less reliable and harder to replicate in English.

the context of the lexical decision task, which does not require grouping stimuli into tightly controlled semantic categories. The following sections report lexical decision experiments that were designed for this purpose.

5.1 Lexical Decision Experiment 1

Three groups of participants completed a lexical decision task in which difficulty of the task was varied by manipulating the orthographic "wordlikeness" of the nonword foils—the tasks were identical in all other respects. The main goal of this experiment was to determine whether substantially increasing the difficulty of the task alone (the hard condition) could result in substantial increases in overall latency and ambiguity effects similar to those found in semantic categorization (i.e., homonymy disadvantage, no difference between polysemous and unambiguous words; Hino et al., 2006). The medium condition was aimed at testing a novel prediction of the simulations presented in the previous chapter, which indicated that during the transition between typical lexical decision and typical semantic categorization results, there should be both a homonymy disadvantage and polysemy advantage relative to unambiguous words. The easy condition was aimed at replicating the classic polysemy advantage in lexical decision (Rodd et al., 2002). The analytical techniques employed to examine the results from first experiment were initially based on standard practices in the field.

Methods

Participants. Students from the undergraduate subject pool at Carnegie Mellon University participated in the experiment in exchange for course credit; 42 participated in the easy condition, 39 in the medium condition, and 40 in the hard condition. All participants had normal or corrected to normal vision and were native English speakers (i.e., English was their first language). Each student participated in only one condition of the experiment.

Apparatus. The experiment was executed on computers running E-prime 1.1.4.1 (Schneider, Eschman, & Zuccolotto, 2010), and was displayed on 17" CRT monitors. Participants responded on a standard keyboard.

Stimuli and design. The experimental word stimuli were taken from Rodd et al. (2002), although to accommodate for dialect differences between British and American participants, two words (CHAP, CRICKET) were replaced with other words (PEER, MAROON) which were matched on word frequency, word length, number of meanings, and number of senses. To briefly reiterate Rodd's design, the word stimuli were chosen so as to vary on both the number of unrelated meanings (one vs. two) to which they were associated, and the number of related senses associated with these meanings (few vs. many). For ease of reference, the words with a single meaning and few senses associated with this meaning correspond to (relatively) unambiguous words (see also Armstrong, Tokowicz, & Plaut, 2012), the words with a single meaning and many related senses associated with this meaning correspond to polysemous words, and the words with two meanings and few related senses associated with these meanings correspond to homonymous words. An additional condition consisting of hybrid words with many meanings and many senses was also included to counterbalance the design but is not discussed in detail here.[2] In addition to the experimental word stimuli, 32 filler word stimuli were generated that were presented at the beginning of each block of trials and during the practice trials. These words were matched on frequency and length to the distribution of frequency and lengths of the experimental word stimuli.

The nonwords used in this experiment were generated by sampling words from the MRC

[2]The population of hybrid items in English is relatively small and as a result, the items in this condition are not as well matched on confounding variables as items in the other conditions. Given this issue and the primary focus of the settling dynamics account on the settling of homonymous, unambiguous, and polysemous words, these items are therefore not discussed in detail in the first lexical decision experiment. Performance for these items is, however, examined in the context of the second experiment, which addresses many of the issues with the first experiment.

database (Coltheart, 1981) and randomly interchanging one consonant with another consonant. The resulting character strings were then screened to ensure that the consonant switching did not produce a word, and that these strings were composed of legal bigrams. In all experiments, nonwords were selected so as to match the distribution of lengths of the word stimuli. For the easy condition, for each string length the positional bigram frequencies of the nonwords were matched to the positional bigram frequencies of Rodd et al.'s (2002) legal nonwords. For the hard condition, for each string length the nonwords with the highest positional bigram frequencies in the set of candidate items were selected. For the medium condition, for each string length the positional bigram frequencies of the nonwords was set to be half way between the positional bigram frequencies for the nonwords used in the easy and hard conditions. The positional bigram frequencies for each string length and condition are listed in Table 5.1.

The orthographic neighborhood of the nonwords in each condition was also compared to that of the experimental word stimuli using Coltheart's N (Coltheart, Davelaar, Jonasson, & Besner, 1977). The neighborhood size of the words (Mean = 7.1, SE = .45) was significantly greater than that of the nonwords in the easy condition (Mean =4.7, SE = .32), $t(286) = 4.2$, $p < .001$, non-significantly different from that of the nonwords in the medium condition (Mean = 7.0, SE = .36), $t(286) < 1$, $p > .05$, and significantly smaller than that of the nonwords in the hard condition (Mean = 10.9, SE = 0.43), $t(286) = 6.1$, $p < .001$. These results provide further evidence that the wordlikeness of the nonword stimuli increases across conditions, and should therefore modulate task difficulty.

Procedure. Participants were instructed that they would be asked to identify whether the groups of letters that appeared on the screen were words or not by pressing the '/' or 'z' with their index fingers on a standard computer keyboard. Word responses were always made with their dominant hand. Participants were asked to respond to each trial as quickly as they could, while also making as few errors as possible. Before beginning the blocks of trials, participants were presented with an example of both a word and a nonword trial, and reminded of which

Table 5.1: Lexical Decision Experiment 1: Word and Nonword Positional Bigram Frequencies

	Condition					
	Easy		Medium		Hard	
N	M	SE	M	SE	M	SE
3	13	.00	20	.00	27	.50
4	110	.22	140	.33	168	1.20
5	253	.16	466	.96	670	6.73
6	612	.53	1291	1.07	1980	14.43
7	991	.45	2376	2.24	3768	37.12
8	1344	.71	2965	2.12	4593	27.58

Note. N = Number of Letters in the nonword; M = Mean; SE = Standard error of the mean.

response keys to press.

The first block was a practice block, consisting of 12 randomly selected filler words and 12 randomly selected nonwords. This was followed by four experimental blocks of trials, interleaved with one minute rests. Each experimental block began with 5 randomly selected filler words and 5 randomly selected nonwords which were not included in later statistical analysis, followed by 32 randomly selected experimental words and 32 randomly selected nonwords. The order of stimulus presentation in each block of trials was random, with the constraint that no more than 3 sequential trials could be of the same stimulus type. In all blocks, each trial began with a fixation cross (+) for 500 ms, followed by the presentation of either a word or nonword character string. The string remained on the screen until the participant responded, or for a maximum of 5000 ms. At the end of each trial the next trial began automatically.

Results

On the basis of the argument outlined by Hino and Lupker (1996), the initial set of analyses reported below focus on generalizing the observed effects across participants and not across items. The motivation for this type of analysis is that between-participant variability is typically

higher than between-item variability. Consequently, item-level analyses, which involve collapsing across data from different participants, will be less powerful and should not influence the interpretation of the results substantially. Rather, generalization across items should be made on the basis of the assumed underpinnings of the relevant effects rather than on the principles of statistical inference. The results of analyses that question these assumptions are described later.

Within-condition accuracy. The overall accuracy for each participant and each word was first screened for outliers relative to the other participants or items in that condition using a z-test and a p-value threshold of .025. In all three conditions, no subject was below 78% accuracy, and all subjects were included in all analysis. No more than three stimuli were eliminated per condition. Descriptive statistics for the accuracy data are presented in Table 5.2, and are depicted for the unambiguous, polysemous, and homonymous words in each difficulty condition in Figure 5.1.

Each difficulty condition was subject to a separate 2 (meaning: one / two) × 2 (senses: few / many) within-subjects ANOVA, and to planned pair-wise comparisons among the unambiguous, polysemous, and homonymous conditions. For ease of comparison, the results of the ANOVAs (both for accuracy and correct latency) are summarized in Table 5.3. All significant effects have p-values less than .05. In the ANOVA of the easy condition data, there were no significant effects. In the ANOVA of the medium condition and the hard condition data, there were main effects of meaning and of sense such that responses were less accurate for words with two meanings and more accurate for words with many senses.

In the pair-wise analyses, all of the conditions showed the same numeric rank ordering, such that homonymous words were responded to least accurately and polysemous words most accurately. These data were also all near ceiling. There were no significant effects in the easy condition (unambiguous vs. polysemous, $t(41) = 1.3$; unambiguous vs. homonymous, $t(41) < 1$; polysemous vs. homonymous, $t(41) = 1.3$). In the medium condition, there were significant differences between the unambiguous and polysemous words ($t(38) = 4.4$), and polysemous and

Table 5.2: Lexical Decision Experiment 1: Accuracy

# meanings / # senses	Condition					
	Easy		Medium		Hard	
	M	SE	M	SE	M	SE
one/few[a]	.97	.006	.97	.004	.97	.008
one/many[b]	.98	.008	.99	.003	.98	.005
two/few[c]	.97	.006	.96	.006	.96	.008
two/many	.98	.005	.97	.007	.97	.006
nonword	.94	.008	.92	.008	.91	.013

Note. M = Mean; SE = Standard error of the mean. [a] i.e., unambiguous words, [b] i.e., polysemous words, [c] i.e., homonymous words

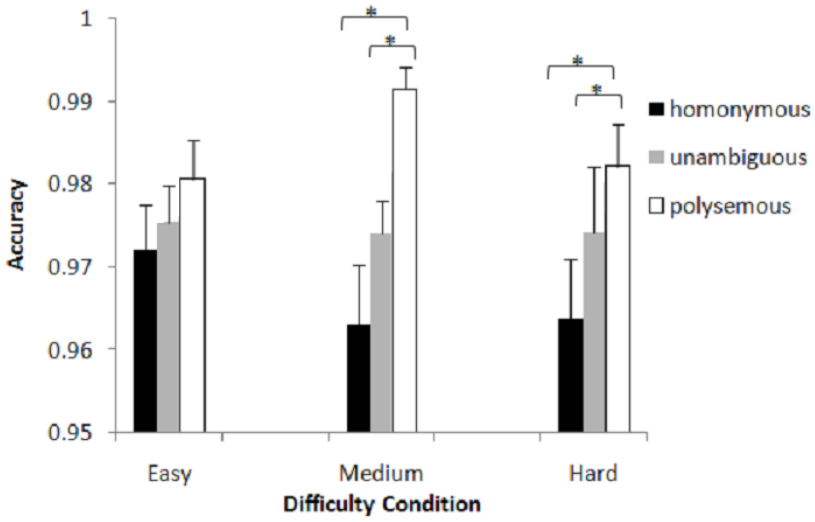


Figure 5.1: Lexical Decision Experiment 1: Accuracies for unambiguous, polysemous, and homonymous words for each difficulty condition. Error bars in this and all subsequent figures depict the standard errors of the means.

homonymous words, ($t(38) = 3.9$), but no differences between homonymous and unambiguous words ($t(38) < 1$). In the hard condition, there were significant differences between the unambiguous and polysemous words, ($t(39) = 2.6$), and the polysemous and homonymous words ($t(39) = 4.0$), but no differences between the unambiguous and homonymous words ($t(39) = 1.2$). Consequently, the evidence provided in this first set of analysis serves primarily to support the presence of a polysemy advantage and a weak homonymy disadvantage in lexical decision. However, the effects in accuracy are generally quite weak, presumably because of performance being near ceiling, which can prevent large differences from emerging, and because of the binomial nature of accuracy data, which limits its sensitivity.

Within-condition latency. Only accurate responses with latencies greater than 200 ms and within 2.5 standard deviations from the mean latency for that level of meaning and sense were included in the analysis; approximately 8% of the trials were dropped for not meeting these criteria (similar results were also obtained with less strict data trimming procedures). Descriptive statistics for the three conditions are presented in Table 5.4, and are depicted for the unambiguous, polysemous, and homonymous words in each difficulty condition in Figure 5.2.

As in the accuracy data, each difficulty condition was subject to a separate 2 (meaning: one / two) x 2 (senses: few / many) within-subjects ANOVA and to planned pair-wise comparisons among the unambiguous, polysemous, and homonymous conditions. The results of the ANOVAs are summarized in Table 5.3; in brief, there were main effects of meaning and of sense in all three difficulty conditions corresponding to slower performance for words with many meanings and faster performance for words with many senses. Additionally, there was an interaction effect in the medium difficulty condition such that the hybrid items did not suffer from as substantial a slow-down as the homonyms despite also having multiple meanings. Similar non-significant interaction trends were observed in the other conditions. Given the potential for a weak interaction effect complicating the interpretation of the main effects, the more sensitive and specific pair-wise comparisons are useful for clarifying these results.

Table 5.3: Lexical Decision Experiment 1: F-statistics for the 2 × 2 Within-subjects ANOVAs

	Condition					
	Easy (df_e = 41)		Medium (df_e = 38)		Hard (df_e = 39)	
Effect	Acc	RT	Acc	RT	Acc	RT
Meaning	< 1	4.3*	5.6*	14.3*	6.0*	13.7*
Sense	3.1	7.3*	11.0*	39.4*	9.7*	10.6*
M x S	< 1	2.5	< 1	5.4*	< 1	2.9

Note. All tests have one degree of freedom treatment. Acc = accuracy. RT = reaction time. df_e = degrees of freedom error. * p < .05.

In the planned pair-wise comparisons, the expected rank ordering of performance was always observed numerically: polysemous words were responded to most rapidly and homonymous words were responded to most slowly. For the easy condition, there were significant differences between both unambiguous and polysemous words ($t(41) = 2.8$), and polysemous and homonymous words ($t(41) = 3.3$), but no significant difference between the unambiguous and homonymous words ($t(41) < 1$). In the medium condition, there were significant differences in all pair-wise comparisons (unambiguous vs. polysemous, $t(38) = 3.5$; unambiguous vs. homonymous, $t(38) = 3.6$; polysemous vs. homonymous, $t(38) = 7.8$). In the hard condition, there were significant differences between unambiguous and homonymous words ($t(39) = 3.8$), and homonymous and polysemous words ($t(39) = 5.5$), but no difference between unambiguous and polysemous words ($t(39) = 1.7$).

Between-condition accuracy. Rounded to two significant digits, the overall accuracy for all of the words in the easy, medium, and hard conditions were all .97 (SE <= 0.001, respectively). Statistical analysis examining differences between these conditions are therefore not reported in detail because all comparisons were non-significant with F-statistics less than 1.

Table 5.4: Lexical Decision Experiment 1: Latency

# meanings / # senses	Condition					
	Easy		Medium		Hard	
	M	SE	M	SE	M	SE
one/few[a]	595	13	585	10	597	9
one/many[b]	579	12	568	10	588	9
two/few[c]	598	12	603	12	619	8
two/many	592	13	574	10	600	9
nonword	695	23	680	15	725	13

Note. M = Mean; SE = Standard error of the mean. [a] i.e., unambiguous words, [b] i.e., polysemous words, [c] i.e., homonymous words

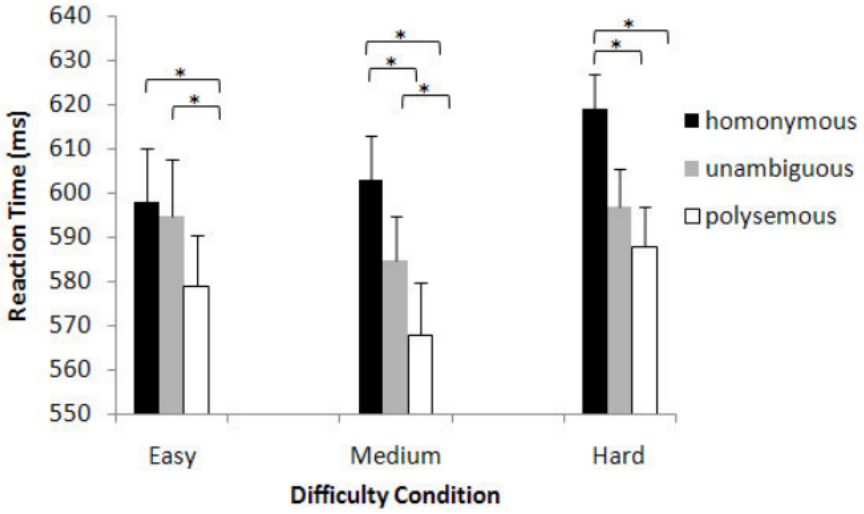


Figure 5.2: Lexical Decision Experiment 1: Latencies for unambiguous, polysemous, and homonymous words for each difficulty condition. Significant differences are indicated with asterisks.

Between-condition latency. The overall latencies for all of the words in the easy, medium, and hard conditions were 591 ms (SE = 1.9), 582 ms (SE = 1.7), and 600 ms (SE = 1.4) respectively. A 3x2x2 mixed factorial ANOVA with one between-condition variable (difficulty: easy / medium / hard) and two within-condition variables (meaning: one / two; sense: few / many) was conducted with the aim of determining whether a main effect of difficulty was present in the latency data; none was found ($F(2, 118) < 1$). Similar follow-up ANOVAs contrasting only the easy-medium, easy-hard, and medium-hard conditions also failed to show any effect main effect of latency (respectively, $F(1, 79) < 1$; $F(1, 80) < 1$; $F(1, 77) = 2.0$ p = .16), although the difference between the medium and hard conditions was marginally significant.

Discussion

At first glance, the results of this experiment mirror the predictions of the settling dynamics account. When decisions are easy and presumably involve a relatively imprecise semantic code there is a polysemy advantage. When decisions are moderately difficult and presumably are influenced by a more precise semantic code, there is both a polysemy advantage and a homonymy disadvantage. When decisions are hard and a presumably a relatively precise semantic code drives the response system, there is a homonymy disadvantage. Additionally, in all cases the trends (both significant and non-significant) were such that polysemous words were always more accurate than unambiguous words, which were in turn more accurate than homonymous words, thus ruling out a potential interpretation of the ambiguity manipulation in terms of a speed-accuracy tradeoff. Unfortunately, the consistency of the results with the theory decreases upon more detailed inspection. The following sections briefly describe several of these issues.

Absence of between-condition differences. Although the within-condition results are in strong agreement with the predictions of the account, the between-condition results are not. Specifically, if increased semantic precision is directly associated with increased overall latencies, significant latency differences (or accuracy differences, in the event of a speed-accuracy

trade-off) between the conditions should have been observed. However, the between-condition differences showed only very weak support for this prediction—accuracies for the word stimuli were stable across the different conditions and although latencies did increase marginally between the medium and hard conditions, there was a nonsignificant increase in latency between the easy and medium condition.

One possible cause of this lack of effect in the between-subject comparisons was that between-participant variability was greater than within-participant variability. Consequently, the failure to find these effects may simply be a result of a lack of statistical power for these comparisons. The nonword data, which were not the initial focus of the planned analysis, did show slightly greater support for the expected predictions—accuracies decreased significantly between the easy and the medium conditions and correct latencies increased significantly between the medium and hard conditions (all pair-wise t-test p-values $< .05$). Optimistically, the lack of between-condition effects may simply have been the artifact of greater within-word versus within-nonword variability cause by the ambiguity manipulation (although other work suggests that the opposite may be the case in general, Joordens et al., 2009), coupled with a small speed-accuracy trade-off. The lack of effects may also have been due to differing degrees of motivation across conditions because different conditions were run at different times during the semester.[3] To evaluate this possibility, two extensions of the first experiment were completed. Because of issues with the experimental items employed in these experiments (discussed in the next section), the results of these experiments are not presented in full detail; however, a brief overview of this work is relevant for motivating the empirical and computational work that is described in later sections and chapters. Both of these experiments were virtually identical to the original experiment with the exception that the difficulty manipulation was a within-participants manipulation rather than a between-participants manipulation and the medium condition was dropped from further study.

[3]Recruitment for all of the conditions was partially interleaved. However, the exploratory investigations related to these conditions, which ultimately became the main experiments, did differ in terms of the number of participants recruited at the beginning vs. at the end of different semesters.

In these experiments the difficulty manipulation occured half way through the experiment, such that for one half of the experiment participants were presented with easy nonword foils and in the other half of the experiment they were presented with hard nonwords (two different groups of participants completed each ordering of the easy and hard nonwords). By employing a within-partipant manipulation, the between-difficulty level effects were expected to be exacerbated (albeit, at the expense of the within-condition effects). To further increase the statistical power of this manipulation, approximately 65 partipants were recruited to participate in each ordering of the easy and hard nonwords (130 total).

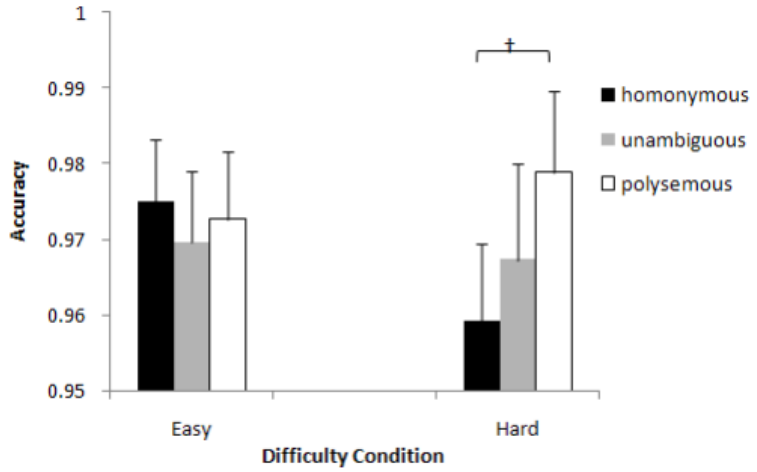
The first experiment failed to show the expected pattern of effects and, in particular, showed 'bleed-over' effects such that the second block typically showed performance typical of the medium condition, as if the expected performance in the second block had been blended with the performance observed in the first block. The second experiment attempted to address the bleed-over issue by introducing a filler task between the two halves of the experiment. This manipulation essentially attempted to 'reset' the participants between the two sections of the experiment so that they began the second half of the lexical decision experiment in a state that was more similar to their state at the start of the first half. This filler task was a pilot version of the numerosity judgment task that is the main focus of Part II of the dissertation. The accuracy and correct latency from this experiment are plotted in Figure 5.3 and Figure 5.4, respectively. The key point to take away from these figures is that the effects observed in each ordering of the easy and hard conditions were not consistent and indicated that the difficulty and order of the two difficulty levels interacted with practice. In particular, presenting the easy block first produced the expected effects in both the easy and hard difficulty conditions; however, presenting the hard difficulty block first caused a homonymy disadvantage in the easy condition during the second half of the experiment. This had several implications for the design of the lexical decision experiment presented next. First, the adaptive effects associated with semantic ambiguity may be ill-suited for study using within-participant designs because of how they interact with the task.

These data are also broadly in line with adaptive effects that have been attributed to how the response selection system engages the lexical semantic system (Armstrong, Joordens, & Plaut, 2009; Joordens et al., 2009), which, if actually the case, would complicate the comparison of the settling dynamics account to the decision making account outlined by Hino et al. (2006). Second, the large effect of practice is likely adding considerable variability to the performance observed in these tasks and should be taken into consideration in the statistical analyses. To this end, all of the remaining analyses have included a number of measures of trial-by-trial performance as additional predictors.

Absence of effects in the item analyses. A second critical limitation of the first lexical decision experiment and the two follow-up experiments is their exclusive focus on between-participant analyses in probing the effects of ambiguity. In addition to generalizing across participants, the generalization of any observed effects to other items is also a critical aspect of psychological—and in particular, psycholinguistic—research. In contrast to the by-participant results, equivalent by-participant analyses failed to find a significant homonymy disadvantage in any of the conditions that were part of the previous experiments (all p-values > .2).

In principle, the lack of equivalent effects in by-participant analyses should cast serious doubt on how the effects of the experiment will generalize. However, several methodological limitations have shifted the perceived importance of these tests. The first of these limitations relates once again to the presence of greater variability within participants relative to within items. A consequence of this is that classic statistical methods that involve averaging across participants to test by-item effects are typically less powerful than analogous by-participant tests that collapse across items. A relatively accepted practice in the field has thus become to ignore the outcome of the by-item analyses when drawing conclusions from an analysis (e.g., Beretta et al., 2005; Hino & Lupker, 1996), let alone attempt to generalize effects to both participants and items simultaneously, which is actually the desired generalization in most cases (Clark, 1973). A second related limitation has been the lack of a method for selecting items that are broadly

EASY BLOCK FIRST:



HARD BLOCK FIRST:

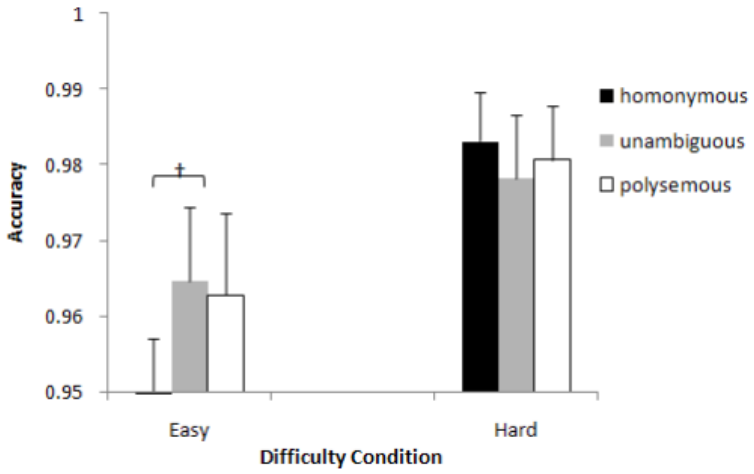
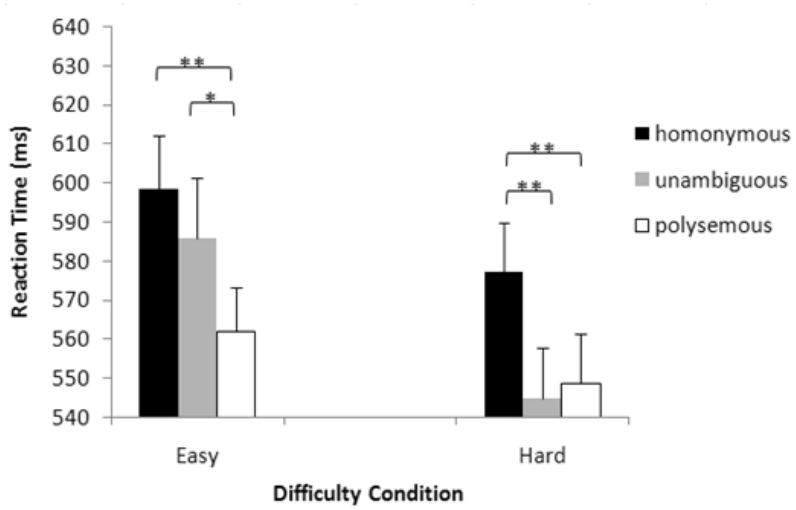


Figure 5.3: Accuracy from the lexical decision experiment that implemented a within-participant difficulty manipulation. Highly significant differences ($p < .005$) are indicated with two asterisks, significant differences ($p < .05$) are indicated with one asterisk, and marginal differences ($p < .15$) are indicated with crosses.

EASY FIRST:



HARD FIRST:

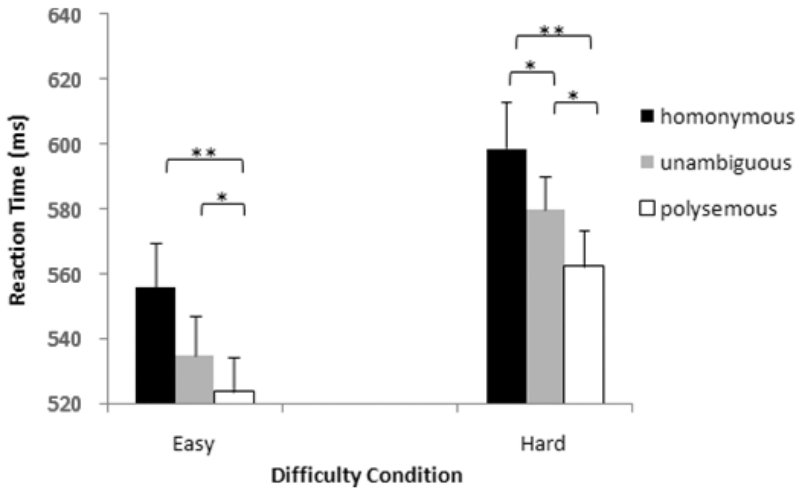


Figure 5.4: Latency for correct responses from the lexical decision experiment that implemented a within-participant difficulty manipulation. Highly significant differences ($p < .005$) are indicated with two asterisks, significant differences ($p < .05$) are indicated with one asterisk, and marginal differences ($p < .15$) are indicated with crosses.

representative of the population they were sampled from (Armstrong, Watson, & Plaut, 2012). Some researchers have even gone to the point of making the unsubstantiated argument that in the absence of such a method and in the context of closely matched item sets, the items used in many studies are in fact not representative of their underlying population (Hino & Lupker, 1996; Hino et al., 2010). Consequently, items are not a random effect (in the statistical sense) that are suitable for broader generalization and are best treated as fixed effects for which by-item analyses would not be necessary.

On the basis of this line of argument, the absence of significant by-item effects of homonymy would not be viewed as a substantial problem by many researchers. However, the view taken in developing this thesis is that adopting such a position and sacrificing the capacity to generalize results to other items ultimately undermines a critical goal of psycholinguistic research. Fortuitously, this is not a unique position in the recent literature and several researchers have attempted to address the first of the two limitations outlined above with improved statistical methods. In particular, Baayen and his colleagues (e.g., Baayen et al., 2008; Baayen & Milin, 2010) have developed a regression technique that allows for knowledge of by-participant and by-item variability to be considered simultaneously as part of a single generalization across both items and participants. This avoids the issue of underpowered by-item analyses due to between-participant variability because the knowledge of participant varability is not discarded when generalizing across items.

Applying the new techniques developed by Baayen et al. (2008) to the data from the previous experiments nevertheless failed to produce a significant homonymy disadvantage in any condition (all p-values > .15). An examination of several of the predictors entered into the analysis indicated that the failure to observe an effect of homonymy was due, at least in part, to the effects of ambiguity being correlated with the effects of other varibles such as frequency and familiarity. These issues were exacerbated further when superior measures of frequency and familiarity were obtained—the former from the SUBTL corpus (Brysbaert & New, 2009), which

has been shown to contain better measures of word frequency than the CELEX data (Baayen, Piepenbrock, & Gulikers, 1995) employed by Rodd et al. (2002), and the latter from a standard familiarity norming study conducted with Carnegie Mellon students (this study was essentially identical to the norming studies reported by Hino et al., 2006). In particular, the different word classes differed both in the mean values of these variables and also on the distribution of these values. For instance, the frequency and familiarity of most homonyms was lower than average and the frequency and familiarity of most polysemes was higher than average.[4] A summary of these confounds is presented in Figure 5.5, which presents the average loadings of the first component of a factor analysis, labeled fLex, on the SUBTL word frequency and the familiarity data associated with the different word classes (similar results were also obtained for each individual variable). Additional analyses focusing on only a subset of the experimental items that were better matched on these variables[5] reduced these differences but the loss of items counteracted any statistical power obtained via this technique. These analyses indicate that the failure to observe a homonymy disadvantage was not due simply to greater between-participant variability but to item-level confounds that are not detectable in by-participant analyses. They further highlight the fundamental problem with ignoring the outcome of by-item analyses, even if the items appear to be relatively well matched on the basis of the summary statistics reported in most papers (cf. Hino et al., 1996; 2010).

No control of meaning dominance. An additional issue discovered with the items used in the first set of experiments was a lack of control for the relative meaning frequency, or dominance, of the different meanings of a homonym. Although this factor has been frequently ignored in much past work, a few recent studies have emphasized the importance of controlling for this

[4]Note that these differences were observed in frequency data for American English and in familiarity norms provided by Carnegie Mellon undergraduates. Consequently, these potential confounds may not apply to the original study conducted by Rodd et al. (2002), which was run in England. However, these issues are expected to have biased the results of the Beretta et al. (2005) replication conducted in Maryland.

[5]This matching was accomplished using an early version of the SOS software reported in Appendix A.

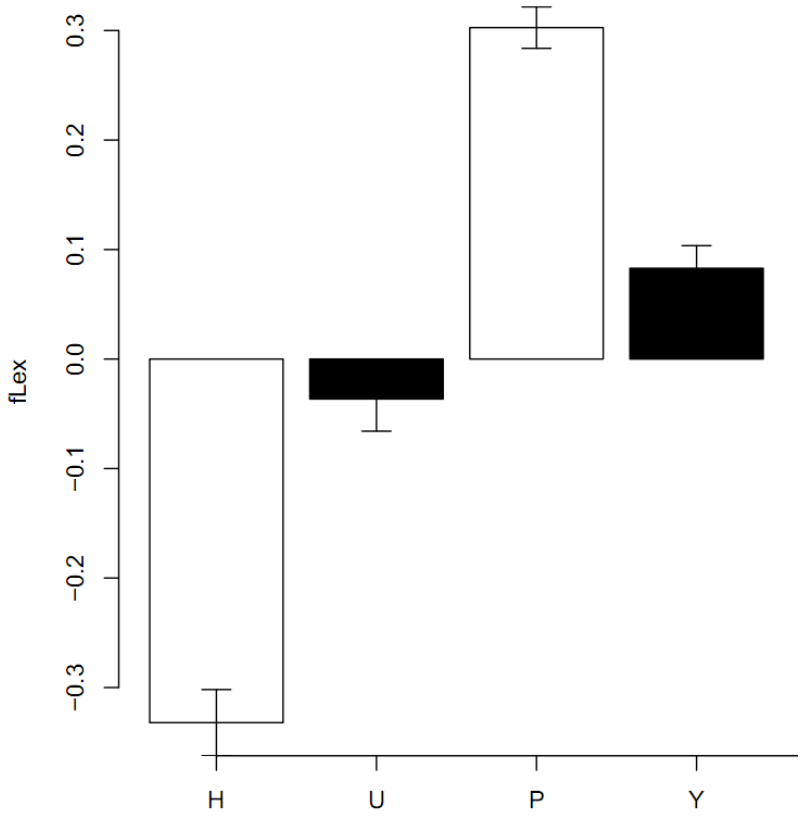


Figure 5.5: Plot of the factor loadings for the first component of a factor analysis, labeled fLex, on the SUBTL word frequency and the familiarity data for the different word classes.

factor (e.g., Klepousniotou & Baum, 2007; Mirman et al., 2010). Essentially, this work argues that it is important to control for dominance because a homonym with one strongly dominant meaning is effectively unambiguous, whereas a homonym with two unequal interpretations is assumed to generate the strongest competition between each distinct interpretation (see also the 'rich get richer' effect, McClelland & Rumelhart, 1981). Preliminary norming of the relative meaning frequencies of the homonyms used in the previous study indicated that a large portion of the homonyms failed to satisfy standard criteria for what constitutes a 'balanced' homonym with relatively equal meaning frequencies (40-60% of the items, depending on the threshold; Klepousniotou & Baum, 2007; Mirman et al., 2010; Twilley et al., 1994). Controlling for this factor post hoc by dropping unbalanced homonyms was not possible given the relatively small set of the items to begin with (only 32 homonyms).

5.2 Summary of Problems and Potential Solutions

In hindsight, it is not surprising that the first set of experiments failed to produce the predicted ambiguity effects. The following section summarizes the problems with this work and how they were addressed as part of the process of designing a new, better study that attempted to provide clearer evidence for a homonymy disadvantage in lexical decision.

Controlling for relative meaning frequency. On the basis of recent research, the homonymous stimuli employed in studies of semantic ambiguity should have approximately equal meaning frequencies to maximize the effects of competition predicted by the settling dynamics account. Identifying these 'balanced' homonyms is a non-trivial undertaking, however, because of the considerable work associated with standard meaning frequency norming techniques (Twilley et al., 1994) and the lack of a recent set of dominance norms for a large set of homonyms, given that these norms change over time (Swinney, 1979). To address this issue, a new norming technique was developed that had participants explicitly rate the relative frequencies of dictionary

definitions using a software tool designed for this purpose. This technique was used to collect meaning frequency norms for 544 homonyms, refered to as the eDom norms, that were suitable for use in standard semantic ambiguity experiments. Appendix B provides the full details of this work and evidence that shows this norming method to produce reliable norms with superior predictive validity than norms produced via other more resource-intensive methods such as rating free associates (Twilley et al., 1994). It also provides evidence related to how the variability associated with this factor can be controlled for either during item selection or during later data analysis.

Selecting a larger, better-controlled item set. A clear limitation of the items used in the original study was the presence of several confounds that may have prevented the detection of a homonymy disadvantage (see also Armstrong, 2007). This issue was exacerbated further by the relatively small size of the item set, which precluded dropping substantial numbers of problem items to address these confounds post hoc. To address these issues, the SOS software package was developed. This software implements a stochastic optimization algorithm that has been tailored for use in the psychological domain, with a particular emphasis on facilitating stimulus selection. In short, this tool automates the item selection process to allow for larger, better-controlled sets of items to be identified than is practical to do by hand. Appendix A provides the full details of this software tool including its formal underpinnings, examples, and evidence that it outperforms manual stimulus selection.

Employing between- instead of within-participant manipulations. An unexpected outcome of the within-participant manipulations were the interactions of the difficulty manipulations with the order of the different difficulty manipulations and with practice. Although these effects may be interesting for other reasons, the most straightforward assessment of a difficuly manipulation appears to be in the context of a between-participant manipulation. This of course raises concerns about the lack of power for the between-participant effects, which can be addressd by

increasing the number of participants in the study.

Generalizing across participants and items. The results of the first experiment highlight the importance of ensuring that observed effects generalize both across items and across participants. Mixed-effect regression models that cross the random effects of item and participant are a recent solution to this problem that also have the advantage of mitigating the effects of between-participant variability on generalizations across items (Baayen et al., 2008).

Powerful difficulty manipulations. Although the lack of between-participant effects may be due, at least in part, to between-participant variability, many other studies have reported large between-participant effects (e.g., Armstrong et al., 2009). Consequently, the lack of effects in the present study may also be due to the use of relatively weak difficulty manipulations. This can be addressed by employing manipulations that have a greater likelihood of substantially altering overall latencies—and by proxy, the predicted ambiguity effects.

5.3 Lexical Decision Experiment 2

The second experiment built on the principles outlined above to design a better assay of the predictions of the settling dynamics account. In particular, the emphasis in this work was on the detection of a homonymy disadvantage in lexical decision, which in retrospect has not been well tested by past work both because of methodological limitations and the non-specific objectives of much of this work (which often focused on showing only different effects for homonyms versus polysemes, and not specific relative performance differences, Rodd et al., 2002).

The most critical update to the previous experiment was the identification of improved methods of slowing overall latencies to tap the later settling dynamics predicted by the account. In particular, the latency differences between the lexical decision studies showing a polysemy advantage and the semantic categorization studies showing a homonymy disadvantage are on the

order of 150 ms (Hino et al., 2006; Rodd et al., 2002). Consequently, the new manipulation should aim to cause substantial changes in between-condition latencies if at all possible. In line with the previous experiment, the first manipulation that aimed to accomplish this was a stronger manipulation of the orthographic "wordlikeness" of the nonword foils and the addition of a pseudohomophone nonword condition, on the basis of previous studies that have shown that pseudohomophones (e.g., TIPE) produce stronger ambiguity effects (Azuma & Van Orden, 1997; Rodd et al., 2002).

The second method that was employed to increase task difficulty deviated more strongly from the earlier experiments and took its inspiration from the results of the auditory lexical decision experiments (e.g., Mirman et al., 2010). In contrast to visual lexical decision experiments, auditory lexical decision experiments have typically reported strong homonymy effects. The auditory tasks are also associated with considerably longer latencies that are close to those reported in semantic categorization tasks (on the order of 800-900 ms). These results suggest that the cause of the stronger homonymy effect is that additional semantic processing has taken place in auditory lexical decision. This additional processing may, in turn, be due to the relatively slow presentation of an auditory word form, which allows for semantic processing to begin early and continue for a longer period of time before the sufficient evidence has been accumulated to generate a response. This causes the auditory lexical decision task to sample from later semantic activation than visual word recognition, in which the full visual orthographic form is available for processing from the outset and orthographic-to-semantic mapping is highly stage-like (Borowsky & Masson, 1996; Borowsky & Besner, 2006).

Still, the most direct test and comparison of the settling dynamics account would be in the context of visual lexical decision, given past literature and the lack of evidence for a homonymy disadvantage in that task. The present experiment therefore attempts to produce a visual analogue of the extended integration of word form information that is a fundamental characeristic of auditory word processing. This was accomplished using a contrast manipulation, which de-

grades the visual word form information. This in turn was predicted to allow for early semantic processing to begin on the basis of the poor initial word form information and consequently allow for the semantic representation of the word to be more fully resolved prior to the availability of sufficient evidence to support a response. Past work using these types of manipulations have also shown that responses are delayed by approximately 100-150 ms, which to a first approximation corresponds to the amount of time needed for processing that may be necessary to observe a homonymy disadvantage. This proposal is not without considerable controversy, however, as some have long-argued for a staged model of orthographic and semantic processing in which orthographic coding is completed first and does not interact with semantic information (e.g., Borowsky & Besner, 1993, 2006). Successfully modulating ambiguity effects using stimulus quality would thus additionally make an important contribution to a more interactive view of orthographic and semantic processing.

Given that both the contrast manipulation and the nonword manipulation may separately contribute to increasing overall response latencies, the experiment investigated whether a homonymy disadvantage would be observed in lexical decision when crossing two between condition manipulations: three levels of nonword difficulty (orthographically hard nonwords vs. orthographically very hard nonwords vs. pseudohomophones) and two levels of stimulus contrast (full [white-on-black] and degraded [grey-on-black]). These six different between-participant manipulations were then crossed with a similar within-participant design to that employed in the first set of experiments: two levels of meaning ambiguity (one meaning vs. many meanings) and two levels of sense ambiguity (few senses vs. many senses).

Methods

Participants. Students from the undergraduate participant pool at the University of Pittsburgh participated in the experiment for course credit. Between 72 and 75 participants completed each condition. Students only participated in a single condition or associated norming study. All had

normal or corrected to normal vision and were native English speakers.

Apparatus. The experiment was presented in a dimly lit room on computers running E-prime 2.0.10.182 (Schneider et al., 2010). Participants responded on a standard keyboard. Full contrast items were presented as white (162.9 cd/m^2) on black (0 cd/m^2), whereas degraded stimuli were presented as dark-grey (1.9 cd/m^2) on black. These values were selected in a pilot study so as to cause at least a 100 ms slow-down by degrading stimulus quality in the "easy" condition.

Stimuli and design. Word stimuli were selected to fill a 2 (meanings: one vs. many) x 2 (senses: few vs. many) factorial design similar to that used by Rodd et al. (2002). For convenience, the one-meaning few-senses cell is refered to as the (relatively) unambiguous condition, the many-meanings few-senses cell as the homonymous condition, the one-meaning many-senses cell as the polysemous condition, and the many-meanings many-senses condition as the hybrid condition. The SOS software package (Armstrong, Watson, & Plaut, 2012) was used to find 100 quadruplets of items (400 total) which were minimally different from one another on a pair-wise level across a number of factors that influence word recognition (see Table 5.5). This minimized the presence of both group- and item-level confounds in the later analysis. Insufficient familiarity, imageability, and meaning frequency data were available *a priori*, so these properties were normed separately[6] with the intent of subsequently discarding problematic items or otherwise

[6]Normative data for familiarity and frequency were collected from approximately 40 different participants. These were standard norming studies similar to those reported by Hino et al. (2006), in which participants rated these measures on a 7-point scale. The only modifications of the standard task that were made consisted of anchoring the extreme ends of the scale using examples of low and high items taken from the MRC database (Coltheart, 1981), and the use of a semi-circular rating scale instead of a linear one. This scale required that participants click in the middle of the circle before the presentation of a probe and then click on the number corresponding to their rating. Although these manipulations appear trivial, together they improved the correlation between familiarity and the raw latency data by 0.2 relative to a standard study that did not involve anchoring the ends of the scale, presented participants with a linear scale, and used the number keys on a keyboard to record responses. Relative meaning frequency was normed by 50 participants using the technique described in Appendix B.

controlling for these factors. Note that this plan was only possible because of the large set of well-matched items identified by the automatic stimulus selection method, which could, in principle allow for over 80% of the items to be discarded due to these confounds and still leave more better-controlled items than have been used in many recently published studies (Hino et al., 2010; Mirman et al., 2010). An additional 100 filler words from the unambiguous population that were matched to the distribution of lengths of the experimental items were selected for use in the practice and warm-up blocks, and at the beginning of each experimental block.

Three different groups of 500 nonwords were generated that matched the distribution of lengths of the word stimuli. Two of these groups were created by sampling from a pool of nonwords created by replacing one consonant in a word in SUBTL (Brysbaert & New, 2009) with another consonant. The easy nonword group consisted of nonwords with positional bigram frequencies that were roughly matched to those of the word stimuli. The hard nonword condition was created by selecting the nonwords with the highest positional bigram frequencies in the pool. Note that in terms of raw values, the easy condition therefore is most similar to the medium condition in the first set of experiments and the hard condition would have been very hard relative to the original findings. This continued use of 'easy' and 'hard' to denote these conditions, despite these changes, is intended to keep the theoretical predictions associated with each level of difficulty consistent with those from the earlier experiments. A third group of pseudohomophones with orthographically existing onsets and bodies and which only contained legal bigrams were sampled from the ARC nonword database (Rastle, Harrington, & Coltheart, 2002). These nonwords were rank ordered based on 1) orthographic Levenshtein distance, 2) orthographic neighborhood size, and 3) positional bigram frequency. The most wordlike nonwords in this list were selected, with the constraint that pseudo-plurals and pseudo-past tenses were largely avoided to prevent participants from basing their lexical decisions on the final letters of the stimuli. Properties of the nonword and word stimuli are presented in Table 5.6.

Table 5.5: Lexical Decision Experiment 2: Properties of Word Stimuli

	unambig.	poly.	homon.	hybrid
example	tango	blind	yard	stall
subtlWF	20.5	21.1	20.8	21.2
length	4.5	4.4	4.4	4.4
num. Meaning	1	1	2.1	2.4
num. Sense	5.6	12.9	6.2	14
wordNet defs.	5.9	12.3	6.7	12.6
posBigram	174.3	192.8	201.3	191.6
N	11.1	11.0	12.3	13.8
LD	1.4	1.3	1.3	1.3
Phonemes	3.6	3.7	3.6	3.7
Syllables	1.2	1.1	1.2	1.1
familiarity	4.9	4.9	4.7	4.7
imageability	4.7	4.8	4.8	4.6
dominance	1*	1*	0.71	0.66
dom. freq.	100*	100*	82	77

Note. Positional bigram frequency and orthographic neighborhood metrics were derived from the SUBTL corpus (Brysbaert & New, 2009). Familiarity, imageability, and meaning frequency were normed after the stimuli were selected and were not matched across quadruplets. *Meaning frequency was assumed to be maximal for these items. subtlWF = word frequency from (Brysbaert & New, 2009). Wordnet defs. = number of definitions in wordNet (Fellbaum, 1998). posBigram = positional bigram frequency. N = Coltheart's N (Coltheart et al., 1977). LD = orthographic Levenshtein distance (Yarkoni, Balota, & Yap, 2008). dominance = [(freq. of dominant meaning - freq. of most frequent subordinate meaning)/freq. of dominant meaning]. dom. freq. = frequency of dominant meaning.

Table 5.6: Lexical Decision Experiment 2: Properties of Nonword and Word Stimuli

	Stimuli											
	Easy NWs			Hard NWs			Pseudo. NWs			Words		
Length	bi	N	LD	bi	N	LD	bi	N	LD	bi	N	LD
3	14	15	1.1	29	31	1.0	25	28	1.0	24	26	1.0
4	121	10	1.4	180	16	1.1	125	15	1.1	125	15	1.1
5	261	4	1.7	608	13	1.3	246	6	1.6	228	6	1.5
6	625	2	1.9	1789	9	1.5	377	4	1.7	603	3	1.8
7	1000	1	2.4	3190	10	1.4	429	1	2.2	766	1	2.1
8	1355	1	2.6	3777	3	1.8	678	0	2.6	806	1	2.3

Note. The word data do not include the filler items. Four and five letter strings made up 85% of the items. bi = positional bigram frequency. N = Coltheart's N. LD = orthographic Levenshtein distance.

Procedure. Participants were instructed to press 'z' or '/' to indicate whether a word or nonword was presented and were provided with examples of each type of trial. Word responses were always made with their dominant hand. To increase the sensitivity of the latency data, avoid speed-accuracy trade-offs, and avoid ceiling effects in accuracy, participants were instructed to respond as quickly as possible and that it was acceptable to make incorrect responses up to 10% of the time. After each block, they were also presented with their latencies and accuracies for that block and for the preceding one. At that point they were instructed to either "try to go faster even if it means making a few more mistakes" if they made less than 10% errors, or to "try to be more accurate, even if it means slowing down a little" otherwise.

The first block was a practice block consisting of 20 trials to familiarize participants with the task, followed by a 100 trial warm-up block to increase participants' proficiency. Participants then completed 8 110-trial experimental blocks, which were seamlessly divided into 10 warm-up trials followed by 100 experimental trials in which the experimental words could be presented.

Only the data from the experimental trials were analyzed. All blocks contained equal numbers of words and nonwords and the order of stimulus presentation was random, with the constraint that no more than 3 trials in a row could consist of only word or nonword stimuli.

Each trial began with a 250 ms blank screen and a fixation stimulus (####+####) presented for a random duration between 750 and 950 ms (the variability in the fixation screen was intended to reduce anticipatory responses and was inspired by a similar manipulation by Purcell et al., 2010). This was followed by a 50 ms blank screen after which a word or nonword stimulus was presented for 4000 ms, or until the participant responded. The contrast of the critical stimulus varied by condition.

Results

Data were screened as follows prior to analysis.[7] All words that at least 10% of participants in the norming studies indicated they did not know and all items with accuracies below 50% were dropped. This eliminated 11 words and 17 nonwords. The eliminated words were distributed approximately equally across conditions. Participants with accuracies below 50% were considered to have failed to adhere to the task instructions and were dropped; this eliminated one participant from the full-hard condition. Next, participants and items were separately screened for outliers in speed-accuracy space using a Mahalanobis distance statistic and a 0.01 p-value cut-off. This eliminated 14 participants in total and no more than three participants per condition. Approximately 4 words were dropped from each of the word conditions, along with approximately 17 nonwords for each difficulty level. Finally, individual trials with latencies lower than 200 ms or higher than 2000 ms, and trial outliers exceeding the z-score associated with $p = 0.005$ within each condition for each block of each participant were dropped (3.4% of trials).

Preliminary data analyses on the full set of items that did not control for the effects of relative meaning frequency failed to find any significant homonymy effects. Consequently, the analyses

[7] Additional analyses using alternative screening methods and cutoffs also produced similar results to those reported here.

that are reported all control for relative meaning frequency in some fashion. The first set of analyses accomplished this by discarding all of the homonym and hybrid items that did not have relatively balanced meaning fequencies, as determined in a meaning frequency norming study similar to that reported in Appendix B. Specifically, items with relative meaning frequenices above 65% were discarded. This cut-off is similar to that employed during item screening in other studies (Klepousniotou & Baum, 2007; Mirman et al., 2010). There were 14 homonyms (mean dominant freq. = 62%) and 22 hybrid items (mean dominant freq. = 59%) that satisfied this constraint. Similar effects were obtained when a 75% cut-off was employed that roughly doubled the number of items in each condition, suggesting a rapid fall-off in the competitive effects across meanings as one meaning begins to dominate.[8] Note that although this screening results in discarding a large portion of the experimental items, the number of remaining items is still comparable to the number of items in many recently published experiments (e.g., Hino et al., 2010; Mirman et al., 2010). These screened data are reported in all of the following tables and figures. A second set of analyses included the data for all of the experimental items but substituted the qualitative division of items as having either one or many meanings with the measure of meaning dominance from the norming experiment. Thus, those analyses attempt to capture how homonyms with relatively balanced meaning frequencies gradually become what are effectively unambiguous words if one meaning of the homonym is strongly dominant. In those analyses, the words with only a single meaning were assigned dominance scores of 100. This made them equivalent to a homonym for which one meaning was intended 100% of the time that the word is encountered. Given the high degree of similarity between the results of the two types of analyses and the non-specificity of the omnibus tests, the results of these analyses are reported only for the planned pair-wise comparisons that are most relevant for evaluating the

[8]Indeed, several connectionist models predict that the degree of competition should vary nonlinearly as a function of dominance (e.g., McClelland & Rumelhart, 1981; Usher & McClelland, 2001). However, given that the exact nature of this nonlinear function was not known, a standard linearity assumption was adopted in the following analyses to keep the analyses simple and avoid overfitting the data.

predictions of the settling dynamics account.

All of the analyses were conducted using a linear mixed-effect model (Baayen et al., 2008) with crossed random effects of participant and item, and as applicable, fixed effects of number of meanings (one / many) or dominance, number of senses (few / many), nonword difficulty, stimulus quality, all of the variables listed in Table 5.5 or some transformation thereof[9] as well as the trial rank, lexicality, accuracy, and latency of the previous trial (based on Baayen & Milin, 2010). Covariates were centered to have a mean of 0. Only meaning, sense, nonword difficulty, contrast, and word frequency were allowed to interact. In the omnibus analyses, positional bigram frequency, Coltheart's N and imageability were not significant predictors of variance and so were dropped from the model. The formal analyses of some overall effects were omitted from this section because of the extemely small standard errors in the data which were approximately two full orders of magnitude smaller than the observed effects; this allowed for reliable inferences to be drawn from the visual presentations of the data. Additional details related to these analyses, including supporting inferential statistics and omnibus analyses, are presented as part of Appendix B and in Armstrong and Plaut (2011). All significant effects have have a p-value less than .05.

Accuracy. Descriptive statistics for the accuracy data are presented in Table 5.7 and in Figure 5.6. The figure also summarizes the results of pairwise statistical tests between the unambiguous word condition and each of the other word conditions. Overall, this figure shows that accuracy levels remained relatively constant across conditions and were slightly above the 90% accuracy threshold that was specified in the instructions and reinforced in the post-block feedback (this is in part due to the data trimming, which tends to eliminate more incorrect responses than correct responses). All of the conditions also showed either a numeric or statistically significant

[9] $\log_{10}(1 + wordfrequency)$ was used instead of raw frequency. Residual familiarity, for which the effects of meaning and sense were first removed, was employed instead of raw familiarity. Raw and residual familiarity correlated strongly (r = 0.98).

trend for the hybrid words to be responded to more rapidly than the unambiguous words; overall, the hybrid items also exhibited performance that was quite similar to the polysemes (consistent with Rodd et al., 2002). This effect diminishes the informativeness of 2 × 2 ANOVAs for each condition that cross meaning and sense because of the interaction between these variables and narrowed the focus of the analyses to the pairwise comparisons. In the full contrast conditions, a polysemy advantage was observed in the absence of a homonymy disadvantage. In the easy-degraded condition, a marginal homonymy disadvantage was observed as well as a marginal polysemy advantage; in the other two degraded conditions a polysemy advantage was observed in the absence of a homonymy disadvantage. The homonymy disadvantage was, however, significant when the continuous measure of dominance was entered into the analysis of the easy-degraded condition (p = .02) and was marginal in the hard-degraded condition (p = .09); the effect was nonsignificant in the pseudohomophone condition (p = .61). No homonymy disadvantages were observed in the full contrast conditions in these additional analyses (ps > .21).

Latency. Descriptive statistics for the latency data from correct trials are presented in Table 5.8 and in Figure 5.7. Overall, these results show small increases in overall latencies as a function of nonword difficulty increases, and substantial overall latency increases as a function of decreases in contrast. These effects are more pronounced for the nonwords than for the words, as in the first lexical decision experiment reported in this chapter. As in the accuracy data, the hybrid items tended to group with the polysemes rather than with the homonyms and were always numerically faster than the unambiguous words, although this advantage was not always statistically significant. In the full contrast conditions a polysemy advantage was always observed and no homonymy disadvantage was observed. In the degraded conditions, a significant homonymy disadvantage was observed in the easy condition but not in the hard or pseudohomophone conditions. In the analyses involving a continuous measure of dominance, a significant homonymy disadvantage was observed in the easy-degraded condition (p = .04), a marginal disadvantage

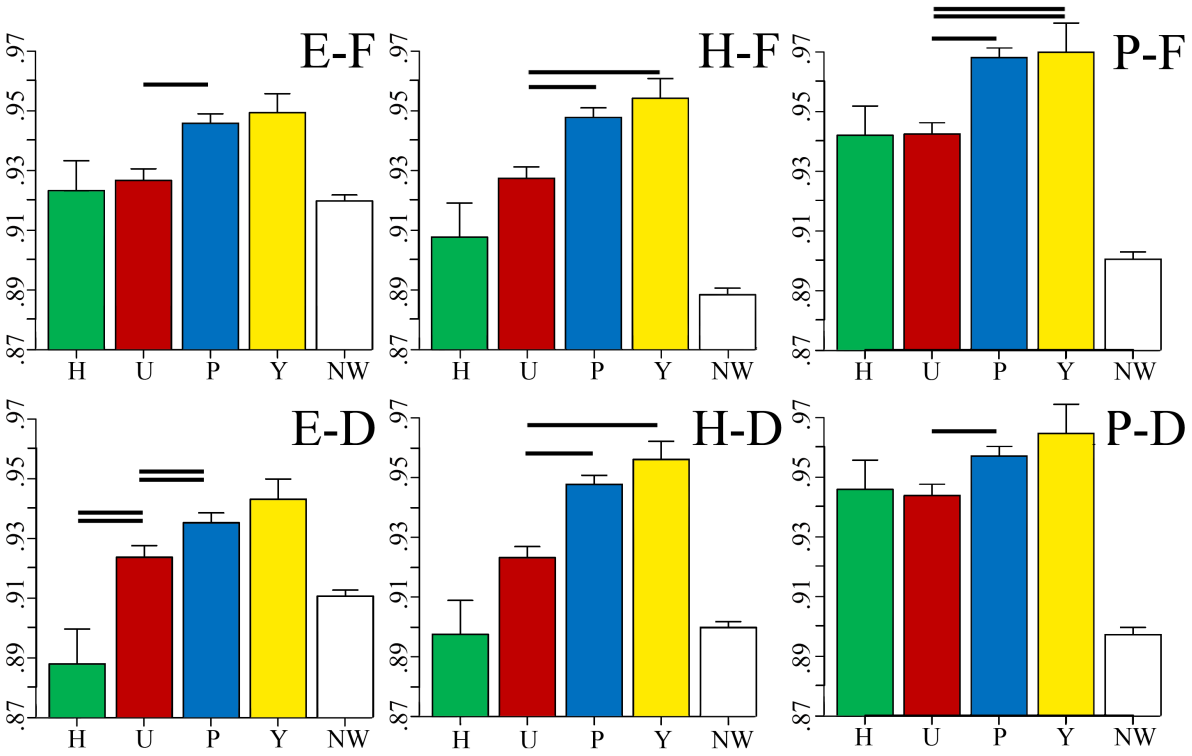


Figure 5.6: Accuracy data for each word class in each condition. E = easy nonwords. E-F = easy-full. H-F = hard-full. E-D = easy-degraded. H-D = hard-degraded. P-F = pseudohomophone-full. H = homonym. U = unambiguous. P = polyseme. Y = hybrid. NW = nonword. Significant ($p < .05$) and marginal ($p < .1$) differences between homonyms, polysemes, and hybrid items relative to unambiguous items are denoted by single and double lines, respectively. Note that these differences were calculated in a multiple regression that included a number of additional covariates. Consequently, the tests are not directly comparing the raw condition means presented here.

Table 5.7: Lexical Decision Experiment 2: Accuracy

	E-F		H-F		P-F		E-D		H-D		P-D	
	RT	SE	RT	SE	RT	SE	RT	SE	RT	SE	RT	SE
homonym	.92	.01	.91	.01	.94	.01	.89	.01	.90	.01	.95	.01
unambiguous	.93	.00	.93	.00	.94	.00	.92	.00	.92	.00	.94	.00
polysemous	.95	.00	.95	.00	.97	.00	.94	.00	.95	.00	.96	.00
hybrid	.95	.01	.96	.01	.97	.00	.94	.01	.96	.01	.97	.01
nonword	.92	.00	.89	.00	.90	.00	.91	.00	.90	.00	.90	.00

E = easy nonwords. H = hard nonwords. P = pseudohomophone nonwords. F = full contrast. D = degraded contrast. acc = accuracy. SE = standard error.

was observed in the hard-degraded condition (p = .15), and no effect was observed in the pseudohomophone condition (p = .51). No marginal or significant homonymy disadvantages were observed in the full contrast conditions (ps > .15).

Discussion

The results of the experiment show that stimulus degradation but not nonword difficulty substantially increased overall latencies. These overall increases in latencies were also associated with the observation of a significant homonymy disadvantage and a slightly weakened sense advantage. Although this result is not consistent with the more extreme shift in expected ambiguity effects that was the aim of this study (i.e., a homonymy disadvantage in the absence of a polysemy advantage), it nevertheless provides empirical support for the settling dynamics account and not the decision system account by showing two different effects predicted by the settling dynamics account within a single task. The tendency for the hybrid items to group more with the polysemes than with the homonyms also suggests that co-operative effects still dominate the overall processing dynamics at this time-point in processing. This provides a more detailed

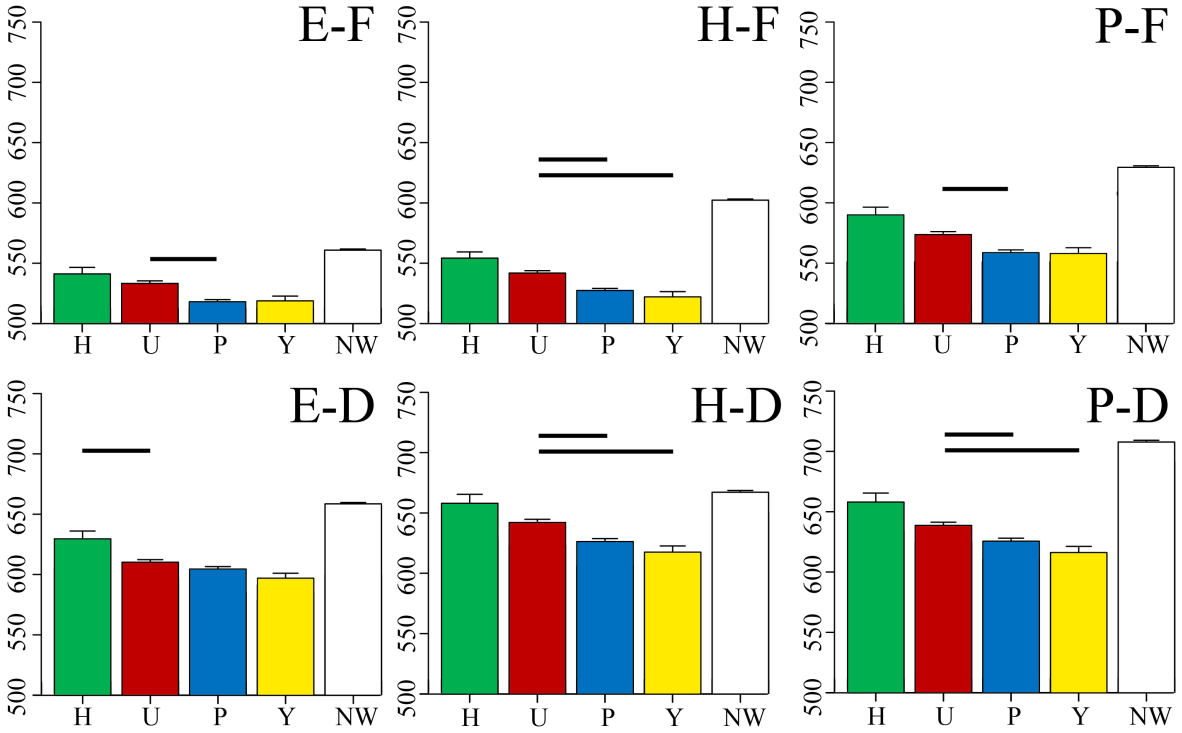


Figure 5.7: Latency data for each word class in each condition. E = easy nonwords. E-F = easy-full. H-F = hard-full. E-D = easy-degraded. H-D = hard-degraded. P-F = pseudohomophone-full. H = homonym. U = unambiguous. P = polyseme. Y = hybrid. NW = nonword. Significant ($p < .05$) and marginal ($p < .1$) differences between homonyms, polysemes, and hybrid items relative to unambiguous items are denoted by single and double lines, respectively. Note that these differences were calculated in a multiple regression that included a number of additional covariates. Consequently, the tests are not directly comparing the raw condition means presented here.

Table 5.8: Lexical Decision Experiment 2: Latency

	E-F		H-F		P-F		E-D		H-D		P-D	
	RT	SE	RT	SE	RT	SE	RT	SE	RT	SE	RT	SE
homonym	542	5	555	5	592	5	631	6	656	6	657	6
unambiguous	530	2	548	2	578	2	612	2	644	2	645	2
polyseme	516	2	535	2	563	2	604	2	628	2	632	2
hybrid	518	3	530	4	561	4	591	4	619	4	620	4
nonword	561	1	603	1	626	2	654	1	660	2	705	1

E = easy nonwords. H = hard nonwords. P = pseudohomophones. F = full contrast. D = degraded contrast. RT = latency (ms). SE = standard error.

constraint for accounts of these phenomena.

The fact that stimulus degradation, in particular, was successful at manipulating semantic ambiguity effects also has important ramifications for models of word recognition more generally. Whereas some researchers argue for separate, non-interactive orthographic and semantic processing stages (Borowsky & Besner, 1993, 2006), the present results support a view of orthographic and semantic processing that involves at least some interaction between those two representations. This is more compatible with the standard processing assumptions made by connectionist models, and suggests that developing theories of semantic ambiguity effects may be a novel and valuable means of advancing theories of word recognition and comprehension more generally.

The failure of the nonword difficulty manipulation also has important ramifications. Although the nonword manipulations failed to substantially slow down overall performance and did not show the predicted ambiguity effects amongst the word classes, responses to nonwords did slow substantially as a function of nonword difficulty. This slowing of only one type of re-

sponse suggests that other aspects of the cognitive system such as the decision system may adapt to the change in stimuli. Indeed, such effects have been predicted and observed in other work that manipulated the perceived accuracy of the nonwords to make them appear more difficult (Armstrong et al., 2009). In that context, adaptation of the decision system alone could account for this type of effect, as was demonstrated via a simulation of an adaptive decision system. This suggests that even within a single task, the decision system may be playing an important role in modulating the observed behavior. An additional related possibility is that the decision system may learn to tap different types of representations to gather the evidence needed to make lexical decisions in the context of different types of nonwords. In particular, pseudohomophones, which are more wordlike overall than the orthographically wordlike nonwords used in the other two conditions, may also engage specific semantic representations associated with the actual word with which they shared a phonological representation. Thus, the informativeness of semantics for making word and nonword decisions will be reduced and the longer latencies in those conditions may be due to the extra time needed to collect evidence for word and nonword responses from non-semantic representations (e.g., orthography, Hino & Lupker, 1996). These results suggest that a similar principle may underlie the results reported by Hino et al. (2010), in which increases in the semantic similarity of the nonwords to the word stimuli were shown to weaken some ambiguity effects despite increasing overall latencies.

5.4 General Discussion

The results of the second experiment, which was designed to address the numerous limitations associated with the first experiment that extended an earlier study by Rodd et al. (2002), generally support the prediction of the settling dynamics account that a homonymy disadvantage will emerge in lexical decision if overall latencies are increased. These results are not, however, perfectly consistent with the simple assumption that increased overall latencies in a task will lead to tapping a later portion of semantic processing to produce different semantic ambiguity

effects; although this may be true to some extent, the characteristics of the stimuli may also interact with the informativeness of the orthographic, phonological, and semantic representations for discriminating between words and nonwords. This in turn may modulate the degree to which semantics influences response generation and impair the detection of a homonymy disadvantage, even if additional semantic processing has taken place relative to another condition. A theoretical approach that is more suited to addressing these complexities is discussed in the next chapter.

Future behavioral work may provide additional insight into these issues. In particular, the results of the present experiments were, at best, consistent with tapping either an early or a moderately precise semantic code, but never reached the precise semantic code that was theorized to be active at the blend state. Additional methodological refinements may enhance some of the weak effects reported here, such as presenting only balanced homonyms from the new set of norms that were collected as part of this work. The failure to produce a homonymy disdvantage in the absence of a polysemy advantage may, however, simply reflect the upper bounds on the difficulty of a standard lexical decision task. In light of the challenges associated with increasing the overall difficulty of a visual lexical decision to show these effects, it is also worth considering how other tasks could provide results that complement the present work. Auditory lexical decision, in particular, may be a useful task for these investigations because previous literature already shows that it can produce a homonymy disadvantage fairly consistently (Mirman et al., 2010; Rodd et al., 2002). This suggests that the auditory version of the task taps a more precise semantic code than visual lexical decision and that making a hard version of an auditory task may show a homonymy disadvantage in the absence of a polysemy advantage. Several different methods may be useful for modulating performance in this task, including the nonword manipulation and analogous stimulus degradation manipulations to those used in the present work, as well as manipulations of the rate at which phonological information is presented over time (i.e., compressions or expansions of the auditory word form), or of the length of the word stimuli, which have been associated with increases in overall semantic activation (Pitt & Samuel, 2006). Insofar

as these studies successfully modulate semantic ambiguity effects, they will provide complementary support for the semantic basis of these effects independent of the specific characteristics of a single task. On the basis of the outcome of the present work, they will likely also contribute to theories visual and auditory word processing, and of the differences and similarities between the systems that underlie these processes.

Chapter 6

Summary and Discussion of Part I

Accounts of semantic ambiguity resolution are challenged by differences in the relative patterns of performance exhibited by polysemous, homonymous, and unambiguous words across tasks. The results of the simulations and experiments reported here provide support for an explanation of many documented ambiguity effects observed across a range of tasks in terms of the settling dynamics of semantic coding. In doing so, the validity and utility of theories of semantic ambiguity effects based on orthography, the decision system, or task-specific factors have also been called into question. Admittedly, this work is not perfect and additional computational and behavioral investigations are needed, as outlined in the previous sections. Neverthless, the evidence presented in Part I is sufficient to support the core principles of the account and show how it can, even in an imperfect form, be used to understand existing findings and guide future research.

Without undermining the value of the current relatively simple account of semantic ambiguity effects, the work described in the previous chapters has also raised an important direction for future work: the need to build more integrated theories of how words are represented and processed, and of how these representations interact with the response selection system to generate responses in particular tasks. This is illustrated at several points in the previous chapters, such as how the manipulation of a visual characteristic of the stimuli lead to semantic effects in the second lexical decision experiment, and how the pseudohomophones showed the slowest overall

latencies but did not show the expected homonymy effects. Additional support for the development of more integrated accounts was provided by the outcome of the computational work, which showed that small differences in how orthography maps onto semantics are contributing to the observed settling dynamics.

Moreover, the development of more integrated accounts would allow for a more rigorous and direct evaluation of the settling dynamics account and of other contrasting accounts. For instance, the present work only indirectly undermined the decision system account proposed by Hino et al. (2006) by showing that contrary to their claims, settling dynamics in semantics are not static and could be a sufficient basis for generating a range of ambiguity effects. However, the present work did not directly compare the contributions of semantic settling dynamics and the contributions of the decision system to generating a response. Rather, it assumed that the decision system could essentially be treated as a constant factor in interpreting the results of the behavioral experiments. This is certainly true to a first approximation relative to how the decision system would need to engage semantics to perform a task such as semantic categorization. However, the absence of effects in the pseudohomophone conditions suggests that a more complex set of interactions between the different representations of a word and of the potential responses in the decision system is taking place. Additional support for this possibility is provided by the contrast between semantic categorization tasks that involve narrow categories, in which no ambiguity effects have been reported, versus broad categories, in which a homonymy disadvantage has been reported (Hino et al., 2006). In both of these cases, additional assumptions, which have yet to be substantiated, are needed to understand how the observed effects fit with the settling dynamics account (e.g., responses in tasks involving narrow categories can be based on a very coarse semantic code). Understanding how the response selection system operates in these tasks and how it taps an ever-changing semantic code would allow for the evaluation of these unsubstantiated assumptions. It would also make explicit how the decision system contributes, or fails to contribute, to generating particular ambiguity effects. Moreover, an explicit computational

instantiation of these principles would allow for the highly interactive and nonlinear dynamics underlying these systems to interact and potentially explain additional effects as an emergent characteristic of this integrated system.

On the basis of this logic, the next major step in the development of a general account of semantic ambiguity effects is to produce a more integrated account of how the stimuli presented in a given task drive the response system to produce different effects. Part II of the dissertation makes some inroads on this front by examining different models of response selection and how they could be adapted for integration with the present work. In doing so, Part II also aims to form the basis for developing more general and wide-reaching theories based upon a single set of learning, representation, and processing principles.

Part II

Towards a General Theory of Response Selection

Chapter 7

Introduction and Brief Review of Computational Models of Response Selection

The next major advance of the settling dynamics account would involve the integration of a model of orthographic, phonological, and semantic processing with a model of response selection. This will permit the modeling of actual tasks and the direct testing and comparison of different theories of semantic ambiguity effects. Obviously, however, understanding how responses are selected in a given task is a generic problem that is relevant to a wide range of tasks beyond semantic ambiguity and has, unsurprisingly, been the subject of considerable prior research. Indeed, this domain represents one of the earlier and richer domains to which computational models have been used to understand and account for performance in a range of different tasks (e.g., Anderson, 1991; Bogacz, Brown, Moehlis, Holmes, & Cohen, 2006; Bogacz, Usher, Zhang, & McClelland, 2007; Gibson, Fichman, & Plaut, 1997; Gomez et al., 2007; Joordens et al., 2009; Link, 1975; Link & Heath, 1975; McClelland, 1993; Movellan & McClelland, 1993; Purcell et al., 2010; Ratcliff, 1978, 1980, 1988; Ratcliff et al., 1999; Ratcliff, Thapar, Gomez, & McKoon, 2004; Stone, 1960; Usher & McClelland, 2001, 2004; Usher, Olami, & McClelland, 2002; Vickers &

Lee, 2000). This chapter examines how this literature can inform the development of a model of response selection that is suitable for integration with the work presented in Part I. The first section reviews the diffusion model and two variants of the connectionist framework that have been employed to understand response selection, so as to identify their strengths and weaknesses. In so doing, this work identifies additional computational principles and supporting empirical data that are needed to develop an improved domain-general connectionist model of response selection which can, amongst many other potential uses, be integrated with the settling dynamics account and its associated simulations.

7.1 Theoretical and Methodological Approaches to Studying Response Selection

An ideal model of response selection should be suitable for (a) modeling a wide range of tasks, (b) modeling the detailed performance observed in a particular task, and (c) understanding the mechanisms underlying reponse selection, and should employ the smallest number of domain-general and independently motivated principles and associated parameters to do so. Obviously, this is a lofty set of objectives and major movements in the field have opted to focus on only a subset of these issues. The following sections review the key strengths and weakensses associated with several major models of response selection and how they relate to these principles.

The diffusion model. The most well-established and well-known model of response selection is the diffusion model (Ratcliff, 1978; Ratcliff & McKoon, 2008), which builds on early accumulator and random walk models of response selection (e.g., Stone, 1960). In essence, the diffusion model is an abstract mathematical model of response selection that simulates this process as the gradual and noisy accumulation of evidence from an initial starting point towards one of two thresholds which, once crossed, signals the initiation of one of the two possible responses. Addi-

tional parameters determine the variability and mean rate of evidence accumulation, the proximity of the starting point of the evidence accumulation process to each of the response thresholds, and other detailed characteristics of the model that can be used to fine-tune its performance.

A key strength of this model is its capacity to fit detailed aspects of the aggregate behavioral data produced in an experiment, including the distributions of errors and correct and incorrect response latencies (Ratcliff & McKoon, 2008; Ratcliff et al., 1999). Indeed, it is exactly this type of evidence that was used to support the conclusion that "It has probably not been realized in the wider scientific community that the class of diffusion models has as near to provided a solution to simple decision making as is possible in behavioral science" (Ratcliff & McKoon, 2008, p. 918). Nevertheless, there are several issues with the model that question the legitimacy of this claim and that motivated the present work to seek an alternative to the diffusion model for simulating response selection.

First, unconstrained parameter fitting algorithms are employed to identify parameters that allow the model to fit a given set of empirical data. Although it has been demonstrated that these algorithms do not allow the model to fit any arbitrary set of data (Ratcliff, 2002), these algorithms nevertheless do not explain how humans learn to perform these tasks and select analogous parameters in their decision systems on a trial-by-trial basis because they are not theories of learning per se. Consequently, these parameter fits serve only to demonstrate that sets of parameters exist that do fit the data, but not to justify the plausibility of those parameters emerging from a principled learning algorithm. There have been some recent efforts to address these issues by developing a learning theory that defines how several parameters are adjusted to optimize performance on a trial-by-trial basis (e.g., Joordens et al., 2009), however, these theories have yet to be expanded to cover the full set of parameters that are part of the diffusion model and in many respects consist of re-inventing well-established error-driven learning algorithms (e.g., Rumelhart, Hinton, & Williams, 1986; Rumelhart et al., 1995).

Second, the diffusion model has been developed as a model of response selection when there

are only two possible choices (i.e., two-alternative forced-choice tasks). Although there have been some efforts to generalize the model to multi-alternative choice tasks (e.g., Krajbich & Rangel, 2011), the degree to which diffusion models can capture behavioral effects has primarily been demonstrated within the context of tasks in which there are only two possible responses. Moreover, the demonstrations of multi-alternative choice have been constrained to tasks in which a fixed and still relatively small set of possible responses is available, and a generalization to tasks with essentially open-ended sets of possible responses has yet to be articulated (e.g., nonword naming, Plaut et al., 1996).

Third, the diffusion model employs a number of parameters and dynamics that were introduced primarily to fit the details of two-alternative forced-choice tasks. The necessity of many of these free parameters is questionable, however, and the data that they fit may simply fall out of alternative formalisms. For instance, variability in the distance from each of the response thresholds is treated as a free parameter to fit aggregate data. However, the aspects of the aggregate data that this parameter is used to fit may simply emerge from a model that simulates performance on a trial-by-trial basis and that allows previous performance to influence future performance.

Fourth, the abstract nature of the diffusion model as a dynamic mathematical equation is rather removed from the neuro-computational mechanisms thought to subserve response selection. Consequently, many recent investigations of the neural bases of decision making (Purcell et al., 2010) can make only loose contact with the diffusion model, although some progress on this front has been made (Ratcliff & McKoon, 2008; Smith & Ratcliff, 2004). Thus, the level of abstraction of the diffusion model fundamentally restricts the degree to which these data can be fit, explained, and predicted within this framework.

Fifth, the diffusion model has typically been employed to simulate simple perceptual tasks that are often borrowed from the signal-detection paradigm (e.g., brightness, Ratcliff & Rouder, 1998; Ratcliff, Thapar, & McKoon, 2003; numerosity, Ratcliff et al., 1999). A key advantage of these simple tasks is that they keep the underlying source of evidence simple. Consequently, the

complex dynamics within the response selection system can be studied in the context of highly-simplified models of the underlying evidence. For example, a standard assumption is that the degree to which a given stimulus supports a particular response is fixed and does not vary as a function of time (i.e., a fixed drift rate in diffusion model terminology, Ratcliff, 1978). Although this type of assumption may be valid to a first approximation in the context of these simple tasks, the same is certainly not true in the case of the settling dynamics account or other models of orthographic-to-semantic mapping (e.g., Armstrong et al., 2009), which show strong nonlinear dynamics in the evidence that is available to support particular types of decisions at different points in time. Again, some early efforts have been made to accommodate changes in the rate at which evidence accumulates and supports a particular response at different points in time (e.g., Ratcliff, 1980), but these attempts have received very little treatment or refinement since they were originally proposed because they are largely unnecessary for studying the standard tasks of interest. Consequently, the degree to which the diffusion model would perform well in the context of a dynamic evidence source is not well established in the literature. Furthermore, this failing highlights the importance of building integrated models such that the model of the evidence that supports different responses and the model of response selection interact. Without such interaction, it is possible that many of the simplifying assumptions each model makes about the other system may be incorrect and invalidate the model.

Sixth, even if the first five issues were addressed, the integration of the diffusion model with the connectionist model of semantic ambiguity described earlier would lead to a heterogeneous model of particular tasks that employs two separate modeling frameworks to model overall performance. On the grounds of parsimony, this is suboptimal because it overcomplicates understanding the detailed aspects of performance within and between the two systems. A better solution would consist of using a single set of learning, representation, and processing principles to simulate the entire system. This highlights the pitfalls of developing a domain-specific model that cannot readily make contact with other systems other than to to treat them as simple ‘black

boxes' that supply a constant amount of evidence that differentially supports one response over another.

Taken together then, despite the diffusion model's ability to fit data with a high degree of precision, this model is lacking in several respects that question its suitability for integration with an explicit instantiation of the settling dynamics account as described in Part I.

Standard connectionist models. In many respects, the performance of connectionist networks that essentially adhere to the standard learning, representation, and processing principles outlined in the 1980's (e.g., Rumelhart & McClelland, 1986a) would appear to directly address many of the limitaitons of the diffusion model, despite not having been developed explicitly as a model of speeded response selection. Theories of error-driven and unsupervised learning provide a principled method of adjusting the weights (parameters) in the system in a constrained fashion that is at least broadly consistent with developmental trajectories (e.g., McClelland & Rogers, 2003). These networks can also adjust their performance on a trial-by-trial basis (i.e., online learning) to simulate the detailed aspects of how systems adapt as a function of experience, provided these adjustments are sufficiently small and a mixture of different types of stimuli are interleaved to avoid catastrophic interference (Rumelhart & McClelland, 1986a). In abstract, the activation of a given unit can be thought of as indicating the degree to which the available evidence (i.e., net input) supports activating a particular representation (localist representation) or feature of a representation (distributed representation). In principle then, it would be reasonable to expect that response selection could be modeled in terms of mapping particular representations from network (e.g., the knowledge represented in the semantic ambiguity networks from Part I) to representations of particular responses. In contrast to the diffusion model, however, this system is in no way constrained to only perform two-choice tasks or tasks in which only a fixed set of specific responses are possible. Rather, connectionist systems can generate responses that were not part of the original training set as a function of the similarity of the inputs to the system to known patterns. Consequently, these systems are suitable for generic response selection prob-

lems in relatively complex tasks such as nonword naming, as has already been demonstrated in previous work (Plaut et al., 1996). In comparison to the abstract nature of the diffusion model, connectionist networks employ neurally-inspired computational mechanisms, which allows these networks to make relatively direct contact with new sources of neural evidence that can inform theories of response selection (e.g., Bogacz et al., 2007; Purcell et al., 2010). Finally, a connectionist model of response selection could be integrated with the existing connectionist model of semantic settling dynamics without any additional modification.

Taken together, a connectionist approach to modeling response selection would appear to satisfy a large number of the desired characteristics of an ideal model of response selection which are not addressed by the diffusion model. There is however, one outstanding issue that would question their applicability for present purposes: the existing literature suggests that these types of models are nevertheless not very good models of response selection. In particular, Ratcliff et al. (1999) compared the results of a simple perceptual judgement experiment to the simulated results produced by two connectionist models—brain-state-in-a-box (BSB; Anderson, 1991) and Graded, Random, and Interactive Networks (GRAIN; McClelland, 1993; see also Movellan & McClelland, 1993)—and by the diffusion model. The authors interpreted their results as indicating that the diffusion model accounted for all aspects of the data, whereas each of the connectionist models failed in important ways. One particularly poignant example of the failure of the connectionist accounts was that they adapted on a trial-by-trial basis using error-correcting learning algorithms such that the accuracy of a given trial was influenced by the accuracy of the previous trial. However, participants did not show any adaptive effects on the basis of their accuracy on previous trials. Additional analyses showed that the connectionist models failed to capture the detailed distributions of accuracy and latency that were fit by the diffusion model. Taken together, these results were interpreted as supporting the diffusion model as the prefered model of response selection, and the possibility of understanding response selection on the basis of the domain general learning, representation, and processing principles of the connectionist

framework was cast into serious doubt. In essence then, the standard connectionist framework would appear to be a near-ideal model of response selection in theory, but appears to suffer the major drawback of not actually modeling the effects that are observed empirically. Even more troubling, these results would appear to undermine the fundamental learning processes that were also employed to model other effects such as those associated with semantic ambiguity. This failure would have broad-reaching implications for the validity of connectionist models in general.

Connectionist models of response selection processes. Deviating from the standard connectionist framework, a modified connectionist framework has recently been developed that specifically aims to capture the detailed effects associated with response selection. This leaky, competing accumulator model (Usher & McClelland, 2001, 2004; Usher et al., 2002) and its variations (e.g., Bogacz et al., 2007; Purcell et al., 2010) (a) employs a threshold-linear as opposed to a sigmoidal input-output transformation function, (b) only allows positive weights into the decision system, (c) explicitly requires the weights between units representing competing responses to be negative, (d) forces units to leak a portion of their excitation, which at minimum must negate any self-excitation that results from a unit sending a weight back to itself, and (e) assumes that unit outputs are noisy but that the amount of noise in their output decreases over time. By virtue of making these assumptions, the parameters of this type of model can be fit in much the same manner as those from the diffusion model to produce fits that are highly similar or slightly superior to those from that model; a rather unsurprising feat given the high degree of similarity—and in some parametrizations, formal equivalence—of the diffusion model and the leaky integrator model (Bogacz et al., 2006, 2007). By virtue of this performance, as well as the model being cast in terms of neural processing mechanics and cross-validating many of the model's properties with those of actual neurons, this model would appear to be an improved model of response selection relative to the diffusion model. This point is reinforced further by the model's capacity to simulate multi-alternative choice problems when there are a fixed number of choices (Usher &

McClelland, 2001) and for fitting data when the degree to which the available evidence supports different responses changes over time (Armstrong et al., 2009; Tsetsos, Usher, & McClelland, 2011).

Despite the leaky accumulator model's capacity to fit data using variations of standard connectionist principles, this model also suffers from many of the same problems that plagued the diffusion model. To date, no comprehensive theory of learning exists to account for how the model's performance improves on the basis of experience with a task because the characteristics of this model are inconsistent with the assumptions of standard learning algorithms. For example, the standard formulation of backpropagation requires that a unit's input-output function is continuous and differentiable for all possible input values. Consequently, this model suffers from the same inability to simulate trial-by-trial adapatation and justify how and why particular parameters emerge during learning that plagued the diffusion model (although some progress on this front has been made, Armstrong et al., 2009). Additionally, to date the representations of the different responses that the model could make have been localist and have been fixed by the modeler; this model is therefore unable to simulate tasks with open-ended sets of responses that are coded as distributed representations. Moreover, although the mechanics of the leaky integrator model are more similar to those of the standard connectionist framework, they still deviate in several major ways. Consequently, developing an integrated model of semantic settling dynamics and of response selection would still require the use of somewhat heterogeneous processing mechanisms. Taken together then, the leaky accumulator model has discarded many of the strengths of the standard connectionist framework to succeed in fitting the details of an aggregated data set with the same degree of precision as the diffusion model.

7.2 Discussion of the Different Approaches and Motivation for the Present Work

Setting aside the increased familiarity that many researchers may have with the diffusion model because of its extended development over the past 30 years, it is clear that connectionist models are superior models of response selection (for most purposes, at least). The main question to answer thus becomes which of the two different types of connectionist models provides the better model of response selection both in general and for the specific purposes of improving the settling dynamics account. In virtually every respect, the standard formalism would appear to be superior, but the literature suggests that it fails to capture empirical phenomena in critical ways that would preclude it as a model of response selection. However, addressing this failing by moving to a leaky accumulator model would entail discarding most of the key strengths of the approach.

Perhaps it is not necessary to choose between these two extreme positions, however. The simulation in Part I that employed additional constraints drawn from the neurobiology literature has already been shown to produce a better recapitulation of the settling dynamics account than the standard approach. A large number of the explicitly-specified or emergent characteristics of that network are also part of the leaky accumulator model (e.g., between layer projections are excitatory, units did not send excitatory weights to themselves, units representing inconsistent semantic features did not send positive weights to one another). It may, therefore, be possible to identify a formalism that achieves a compromise between retaining the domain-general principles of standard connectionist models and the more detailed quantitative fits of the leaky accumulator model, all while increasing the degree to which the model agrees with existing knowledge of neurobiology.

Fundamentally, however, this work would appear to be futile if the conclusions drawn by Ratcliff et al. (1999) regarding the lack of evidence of error correcting learning in behavioral

data are incorrect. Intuitively, these results would not appear to agree with how participants learn to perform tasks—if participants do not attend to evidence regarding the correctness of their response and make adjustments to minimize errors, how can they ever expect to become proficient at many tasks? This intuition is further supported by the results of a lexical decision experiment in which participants clearly attended to (potentially false) feedback to adjust their performance to minimize either actual or apparent errors (Armstrong et al., 2009). Consequently, it is important to identify why Ratcliff et al.'s results do not agree with the results of the Armstrong et al. study and in so doing evaluate the generality of error-driven learning and other characteristics of these models. Additionally, insofar as the results of the original experiment are called into question, this would also necessitate the collection of new baseline data against which connectionist (and other) models could be compared to better contrast their respective strengths and weaknesses. The next chapter aims to accomplish these goals and in so doing inform the development of an improved connectionist formalism suitable for simulating a wide range of phenomena including—but not limited to—response selection, which will be presented in a later chapter.

Chapter 8

Re-evaluating the Empirical Evidence Used to Assess Models of Response Selection: A Behavioral Investigation of Numerosity Judgment

To date, the central tendency of the decision making literature has been to focus on modeling the results of one broad class of task—those tasks involving two-alternative forced-choice responses, particularly from simple signal detection paradigms —thus partially underweighting the breadth of the target phenomena to emphasize the depth and detailed model performance within this class of task. The results of these experiments have in turn been used to develop and discriminate between a wide range of models of response selection, typically by comparing how well different accounts are able to fit the data and by prefering models that produce better quantitative fits (e.g., Stone, 1960; Link & Heath, 1975). Exactly this type of approach was adopted by Ratcliff et al. (1999) to compare how well connectionist and diffusion models of response selection captured the results of a simple perceptual judgment task, which ultimately led to their conclusions regarding the deficiencies of connectionist models. Insofar as the characteristics of

this task were atypical, however, the original conclusions drawn by Ratcliff et al. may be inaccurate regarding the general suitability of connectionist models to account for phenomena related to response selection. This chapter aims to assess the validity of the behavioral results reported by Ratcliff et al. as a 'gold standard' for model comparison. This is accomplished first by considering whether certain esoteric aspects of their investigations may have led to atypical behavioral results that question the fundamental assumptions of connectionist models (e.g., commitment to error-driven learning). This outcome of this evaluation motivated an extension of their original study to determine whether minor changes in the characteristics of the task that make it more representative of standard experiments leads to major alterations in participants' performance that is more in line with the predictions of connectionist accounts. The implications of this work with respect to the evaluation and development of connectionist (and other) models of response selection are discussed in the final section.

8.1 The Ratcliff et al. (1999) Numerosity Judgement Experiment and Associated Simulations

To evaluate and contrast the performance of the diffusion model to the connectionist models, Ratcliff et al. (1999) designed and ran a numerosity judgement experiment to serve as a gold standard for model comparison. This task has been employed in various forms in several past experiments (for brief reviews, see, e.g., Espinoza-Varas & Watson, 1994; Ratcliff et al., 1999) and, in general terms, produces results that are broadly similar to many other simple signal detection tasks and tasks that arguably involve more complex sources of evidence, such as lexical decision (Ratcliff, Gomez, & McKoon, 2004; see Ratcliff et al., 1999, for additional discussion). Ratcliff and colleagues therefore opted to tailor a version of this task for their own investigations.

In brief, this task involved presenting participants with a 10×10 array that could be filled with 0–99 asterisks. The number of asterisks presented on a given trial was determined first by

choosing to sample from a distribution which contains either a ‘low’ or ‘high’ mean number of asterisks, and then sampling from this distribution to determine the specific number of asterisks to present on the trial. Critically, the low and high distributions overlapped substantially, such that on any given trial either of the two underlying distributions could have produced the observed number of asterisks, although one distribution was more likely to have done so on average (except at the cross-over point between the distributions). Participants were tasked with making a response indicating whether the displayed number of asterisks was produced by either the high or low distribution, and were explicitly instructed that on a given trial, the presented number of asterisks could have come from either distribution. To study how the response system adapted to environmental change, Ratcliff et al. (1999, Experiment 2) also varied the likelihood of sampling from either the high or the low distribution across different blocks of trials such that in some blocks, 80% of the presented stimuli came from either the low or the high distribution. Additionally, to maximize the amount of data collected from individual participants and to model between-participant differences, each participant in these experiments completed approximately 12,000 trials in multiple sessions spread across several weeks. The connectionist models were trained on approximations of this same task and the diffusion model was fit to the behavioral results obtained from this task to evaluate how well each model faired in replicating the human patterns of responses.

Ratcliff et al. (1999) noted several disagreements between the data produced by the human participants and those from the connectionist models, of which the most relevant for the present work are briefly reported here. First, the connectionist models trained with error-correcting learning algorithms showed that the accuracy of the previous trial influenced performance on subsequent trials, whereas this was not true in the human data. Second, the connectionist models failed to adapt to changes in the distributions in the same number of trials as observed in the human participants—the models tended to adapt far more slowly. Taken together, these results present a paradox for connectionist models: in one respect, the models appear to be influenced

by knowledge of the accuracy of previous trials although the participants were not, whereas in the other, the participants appeared to be more strongly influenced by the accuracy of previous trials than the models. Insofar as these findings are representative of a broad set of tasks, these results would appear to question the fundamental learning mechanisms underlying the majority of connectionist models, and by proxy the models themselves, which generally make strong commitments to the importance of learning (e.g., McClelland, McNaughton, & O'Reilly, 1995; McClelland & Rogers, 2003; Rumelhart & McClelland, 1986a).

Potential limitations of the study. There are, however, several potential limitations with the Ratcliff et al. (1999) study that restrict the weight that should be placed on these conclusions. First, the participants in this study received extensive practice (12,000 trials), which may have allowed them to reach an asymptotic level of performance not observed in standard experiments, because these typically employ only one tenth as many trials at most. In and of itself, this questions the representativeness of these data of typical performance but may also interact with a second issue: participants were given explicit instructions that the two distributions of items overlapped. This knowledge, coupled with their extensive practice, may have led participants to show atypical adaptation following an incorrect response. For instance, given that participants learned to perform very well on this task, the number of errors that they made was substantially reduced, which is predicted to lead to a corresponding reduction in the number and/or magnitude of the error corrections that result from an error-driven learning algorithm (Ackley, Hinton, & Sejnowski, 1985; Rumelhart, Hinton, & Williams, 1986; Rumelhart et al., 1995). Moreover, participants may have come to know the statistically optimal pattern of responses and to expect that these responses would nevertheless lead to incorrect responses on a subset of trials. Consequently, the fact that participants' perfomance did not differ following an incorrect response may not be representative of performance in many other tasks in which participants have no explicit knowledge regarding the accuracy of the feedback or the overlap in the stimulus distributions, nor in which do they have extensive practice. Indeed, data from a lexical decision experiment

in which false feedback was provided to participants showed that participants will in fact adjust their performance on the basis of perceived errors (Armstrong et al., 2009). Rather, these particular characteristics may cause an expected quantitative reduction in the adjustments that are applied following incorrect responses if the learning rate self-adjusts to a smaller value to compensate for the reduced reliability of the error signal on a given trial (see, e.g., Hinton, 1989; Jacobs, 1988).

Additionally, these properties of the task may cause qualitative changes in the strategies that participants employ when learning to perform the task (Botvinick, Braver, Barch, Carter, & Cohen, 2001; MacDonald, Cohen, Stenger, & Carter, 2000). Consequently, the degree to which the connectionist models, which as they were implemented were explicity required to 'believe' the accuracy of the feedback and to adapt to it on a trial-by-trial basis, failed to capture this pattern of results may not be indicative of a general failure of error-correction mechanisms. A third point, related to the second, is that participants in the second experiment reported by Ratcliff et al. were again given extensive practice with the fact that the likelihood of sampling from one of the two distributions would certainly change at particular points during the experiment. Such changes have already been documented in past literature as being more appropriately modeled by the introduction of a 'task-switching' representation (or more appropriate for the present context, a distribution-switching representation) which was not included in the simulations conducted by Ratcliff and colleagues; instead, those simulations only allowed the models to represent and learn task information using simple mechanisms that are not well-suited to deriving an internal representation of this task structure. Those simulations may, however, be well suited to capturing earlier performance that may be more sensitive to the base statistics of the task environment prior to the development of these additional strategies. Regrettably, the data from the first sessions of the Ratcliff et al. (1999) experiments (1,200–2,400 trials), which are most representative of the length of standard psychological experiments and which would be least likely to show such strategic distribution-switching effects were excluded from the reported analyses.

Discussion of additional potential limitations specific to the methods used to instantiate the connectionist models are deferred to the next chapter, but the factors outlined above clearly indicate that the task employed by Ratcliff et al. (1999) is sufficiently atypical for it to lose its face validity as a broadly representative task for model comparison. Rather, these issues may be more suitably examined in an alternative task in which participants' earlier performance (i.e., during the first 1,200 trials) provide the target phenomena for modeling and in which participants are not provided either with experience or with explicit instruction that could lead to strategic changes in how they perform the task—information that is typically not provided to simple models of response selection. This next section reports an extension of the original Ratcliff et al. study that has been modified to better suit these objectives. Insofar as the results of this study diverge from the results reported in the original study, this will serve to motivate a revision of the conclusions drawn by Ratcliff and colleagues regarding the suitability of connectionist models for simulating response selection, and also serve as a revised empirical target for which connectionist (and other) models should account.

8.2 Numerosity Judgment Experiment

In essence, the experiment reported here is an extension of the first and second experiments reported in Ratcliff et al. (1999). The main goal of the experiment was to examine the generality of the findings reported by Ratcliff and colleagues under a slightly different set of conditions, notably: (a) when participants are initially learning to perform a task and are moving towards asymptotic levels of performance, (b) when external feedback is accurate and can, in principle, provide a basis for learning to respond perfectly accurately on every trial, and (c) when participants have no or minimal prior expectancies or experience with changes in the distributions of the stimuli.

Note that one indirect consequence of running this extension of the original study is that to obtain a similarly large number of total trials, it was necessary to run a larger number of

participants on a smaller number of trials. The following results therefore focus on the group-level trends and make no attempt to examine individual differences because of the relatively small amount of data available from each participant. Nevertheless, a logical future step for this work would be to examine the individual differences across participants. These data could then be used to constrain models of these effects even further by ensuring that the models are sufficiently flexible to account for individual patterns of data without being underconstrained and able to fit arbitrary and contrived data sets (Ratcliff, 2002). For present purposes, this is left as a target for future research, although the data from the individual participants in this study may serve as a basis for some of this work.

Methods

Participants. A total of 121 Carnegie Mellon undergraduates participated in the experiment in exchange for course credit. One participant was excluded because their age (68 years) differed substantially from the mean age of the participants (20 years). Of the remaining participants, 72 were male, the mean age was 20, and ages ranged from 18 to 23. All of the participants were right handed[1] and had normal or corrected to normal vision.

Apparatus. The experiment was implemented using PsychoPy 1.71.01 (Peirce, 2007) and was displayed on 43 cm (17") CRT monitors running at 85 Hz. Asterisks were displayed in a 10 × 10 grid that appeared in the center of the monitor. The grid was approximately square with a length of 13 cm. Participants were seated approximately 55 cm from the screen. Responses were recorded on standard computer keyboard using the 'z' and '/' keys.

[1] Initially, no explicit screening for handedness was employed. However, when the experiment neared completion only seven left handed participants had been recruited. To simplify the later analyses and avoid interactions between the speed of responses made using the dominant hand and the other manipulations in the experiment, the data from these participants were discarded and replaced, to the extent possible given constraints on recruitment, with data from right handed individuals.

Stimuli and design. The stimuli used in the experiment consisted of presenting different numbers of asterisks in the 10×10 grid. These stimuli were divided into 'low' and 'high' categories on the basis of whether the number of asterisks fell above or below a threshold value. For the experimental stimuli, the number of asterisks was sampled from a trimmed normal distribution with a mean of 50, a standard deviation of 14, and upper and lower bounds of 28 and 72, respectively. These bounds were introduced so as to include examples of relatively low frequency items from the extreme ends of the distributions, which were not well characterized by the connectionist models reported in Ratcliff et al. (1999), while at the same time ensuring that these low frequency items occured with a sufficient frequency so as to permit statistical analyses that were reasonably reliable.[2] Depending on the experimental condition, the threshold that delineated a 'low' response from a 'high' response could be either 4.5 points below (i.e., 45.5) or above (i.e., 54.5) the mean of the distribution. This full set of parameters was selected during pilot experimentation to ensure that the judgements were difficult and participants were not immediately at ceiling, that the observed effects were relatively large, and that changing from one threshold value to the other was not overtly detected by most of the participants, and thus increase the likelihood that they would invoke a more sophisticated adaptation strategy. In general, however, the distributional characteristics of the stimuli are similar to those employed in the Ratcliff et al. (1999) experiment, with the notable difference being that the two distributions of stimuli did not overlap.

In total, participants were presented with three mega-blocks of 400 experimental stimuli. The threshold for making 'low' or 'high' responses could potentially differ across each of these

[2]Note that a consequence of varying response frequency as a function of distance from the mean of the distribution is that these two variables are highly correlated ($r = .99$), and so the independent effects of each of these variables cannot be assessed. Consequently, the later analyses that involve a measure of the distance from the threshold can also be interpreted as resulting from the frequency with which that stimulus was presented. This distinction is not important for present purposes, but follow-up studies that vary distance and frequency independently may be useful for evaluating the fine-grained predictions of models with respect to these variables.

blocks. Given the number of participants in the present study this generated more data overall than in the Ratcliff et al. (1999) study, although each individual participant generated only about one tenth as many trials. The full set of 16 different combinations of thresholds, order of thresholds, and response key combinations were run so that data from 7±1 participants was included in each of these conditions after the statistical screening. These trials were presented across 25 experimental blocks each containing 50 trials except for the first and last blocks, which contained 25 trials. The shorter duration of the first and last blocks ensured that the transition between mega-blocks and associated thresholds occured mid-way through a block of 50 experimental trials and not after a self-paced break between blocks. This prevented the effects of adaptation from being confounded with the tendency for participants to respond more slowly and/or less accurately at the beginning of a new block, which has been observed in previous work (Armstrong & Plaut, 2008, 2011).

The generation of the stimuli in each mega-block began by calculating the probability density function for every number of asterisks in the [28,72] interval. The values in this density function were then multiplied by the total number of trials in the mega-block and rounded to the nearest integer to determine the frequency with which each number of asterisks would be presented (one trial was removed from the number of trials containing 50 asterisks to keep the distribution symmetric around the mean for the desired number of trials). This method was prefered over a strictly probablistic sampling method (like the one used in the original study) because it eliminated variability in the distributions of asterisks presented across participants and across mega-blocks within participants. These stimuli were then divided into 'low' and 'high' numbers of asterisks on the basis of the response threshold for the mega-block. Given the threshold levels used in the experiment, 72% (36% × 2) of trials fell on the two tails of the distribution and were always either low or high, respectively, across participants. The remaining 28% of trials near the center of the distribution could be assigned to either the low or high condition depending on the threshold employed in a given mega-block. The frequency of each type of response was thus

similar but slightly less extreme than the 80:20 ratio employed by Ratcliff et al. (1999) when varying the probability of sampling from different distributions.

Additionally, participants were presented with a single block of 10 practice trials before beginning the experimental trials. These practice trials were uniformly distributed and spaced across the range falling within one standard deviation of the mean of the distribution, and employed the same threshold level that was to be used in the first block of experimental trials for a given participant. This distribution of practice trials ensured that participants saw seven items from one response category and three from the other, which approximated the likelihood of each of these types of responses in the main experiment. The use of a narrower set of upper and lower bounds in the practice stimuli was intended to approximate the greater likelihood of observing stimuli in this range during the experiment. The small number of practice stimuli was employed so that the practice block would primarily serve to ensure that participants understood the task and to allow for early adaptive effects to be examined in the experimental trials. This contrasts with and complements the methods employed by Ratcliff et al. (1999), who discarded the data from the first 1,200–2,400 trials in their experiments.

The order in which the practice trials and experimental trials were presented was randomized for each participant with the constraint that no more than five trials of a same type could be presented successively within a mega-block or practice block. Additionally, the positioning of the asterisks within the array was randomized on each trial.

Procedure. Participants were instructed that they would be presented with displays containing different numbers of asterisks and would have to decide whether the number of presented asterisks on a given screen was either ‘low’ or ‘high.’ They would have to learn what constituted a ‘low’ or ‘high’ number of asterisks by making responses and learning from the feedback provided after each incorrect response and at the end of each block of trials. Note that in contrast to Ratcliff et al. (1999), participants were not instructed that a given number of asterisks could be produced by either the ‘low’ or ‘high’ distributions (because the feedback in the present ex-

periment was accurate within each mega-block), nor were they informed that the threshold that delineated between the 'low' and 'high' distribution might change during the experiment. This decision was intended to increase the likelihood that participants would adapt to the changes in the characteristics of the stimuli using implicit learning mechanisms, rather than invoking more sophisticated mechanisms for adapting to anticipated and substantial shifts in these characteristics (Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999).

Participants responded by pressing the 'z' and '/' keys with their left and right index fingers. The type of response associated with each of these keys was counterbalanced across all of the experimental conditions between participants. Participants were instructed to respond as quickly as they could without making many mistakes. To operationalize this instruction, instead of using a payoff scheme as in Ratcliff et al. (1999), participants were instructed that after each block of trials they would receive a message instructing them either to "try to go faster, even if it means making a few more errors", or to "try to make a few less errors even if it means slowing down." The accuracy threshold that determined which of these two messages was to be presented was 90%, although this was not known to participants. In previous studies, this type of instruction has been found to encourage efficient (i.e., fast and relatively accurate) responding and also facilitates the detection of effects in both the latency and the accuracy data (Armstrong et al., 2009; Armstrong & Plaut, 2011). This procedure also has the advantage of increasing the number of what are often relatively scarce error trials, which have, in some instances, played an important role in evaluating theories of decision making (e.g., Ratcliff et al., 1999; Stone, 1960), and which were arguably ill-defined in the Ratcliff et al. study.[3] As part of the instructions, participants were presented with an example trial of both a 'low' and a 'high' number of asterisks sampled near the lower and upper bounds of these two distributions, respectively.

[3]The Ratcliff et al. (1999) 'error' feedback was ill-defined in the sense that participants could be told they made an error despite making what was probablistically the correct choice. Although the present work does avoid this issue within mega-blocks it does not, admittedly, eliminate this problem entirely because of potential changes in the threshold between megablocks.

Following the instructions, participants were presented with the short block of practice trials followed by feedback on their performance in the block. Subsequently, they were asked if they required any clarifications about the experiment. Then they began the 25 experimental blocks, each of which ended with a screen providing them with feedback on how they should try to respond in the next block and an opportunity to take a self-paced break.

Each trial in the experiment consisted of the following, presented in white text on a black background: (1) a fixation stimulus (+) which appeared for 500–700 ms (the variability in the fixation duration was motivated by a study conducted by Purcell et al., 2010, which suggested that such manipulations, at least for high levels of variability, reduce anticipatory effects), (2) a blank screen for 50 ms, (3) the 10 × 10 array populated with a number of asterisks, which remained on the screen until participants responded or for a maximum of 5000 ms, and (4a) if the response was incorrect, "INCORRECT" appeared on the screen for 400 ms, (4b) if no response was made within 5000 ms, the text "Too slow. Please respond more quickly" was displayed for 2000 ms, (4c) if participants pressed a key other than 'z' or '/', "INVALID RESPONSE. Response keys are 'z' and '/' " appeared on the screen for 2000 ms, or (4d) if participants responded correctly, the next trial began automatically.

Following the completion of the experiment, participants were debriefed and asked if they thought that the threshold had changed at specific points in the experiment. The duration of the entire procedure was approximately 45 minutes.

Results

Data screening. The results of the experiment were analyzed using R (R Development Core Team, 2011). Prior to the main analyses, participants were screened to confirm that none had accuracies below 50%, which could indicate a lack of adherance to the task instructions. Next, participants in the conditions employing a constant threshold (same–same), a threshold change after the first block (change–same), a threshold change after the second block (same–change),

and a threshold change after both the first and the second block (change–change) were screened for atypical overall performance; simultaneously, a similar screening was applied to the full set of participants on the basis of their performance in the first mega-block. The latter analysis was aimed at minimizing between-participant variability that could influence mean performance across the conditions. These screenings eliminated participants with atypical combinations of speed (for correct responses) and accuracy, and were accomplished using Mahalanobis distance statistics with two degrees of freedom tested against a one-tailed χ^2 distribution with a critical p-value of 0.01. A total of seven participants were eliminated for exceeding these criteria, with between one and three participants being eliminated from each individual condition. This left 28±2 participants in each of the conditions. Next, individual trials from the experimental blocks were screened to remove invalid button presses and trials for which no response was made before the automatic timeout (0.1% of trials), and trials with latencies below 200 ms and above 2000 ms (1.0% of trials). Unless otherwise noted, all of the following analyses were conducted using these screened data from the experimental blocks, or a subset thereof.

Initial proficiency. To determine the degree to which participants had successfully pre-configured their response selection system prior to beginning the experiment, the first analysis focused on the data from the very first practice trial for all of the participants. A binomial test (Clopper & Pearson, 1934) showed that the accuracy on this first trial was 63% (SE = 0.05), which was significantly higher than that expected by chance (50%), $p = 0.008$, $n = 104$. This suggests that prior to beginning even the practice trials, participants have used the task instructions and their existing knowledge of numerosity to configure their response system in a way that is generally in line with the task, without having any experience performing the task itself. A similar analysis on the first experimental trial indicated that accuracy had already reached 82% (SE = 0.04) on the basis of the experience with the task that was provided in the 10-trial practice block, $p < 0.001$, $n = 112$.

Overall performance in the constant threshold conditions. The data from the participants for which the threshold was constant throughout all three mega-blocks were examined to evaluate the general characteristics of the responses in the experiment in the absence of any adaptive effects resulting from changes in the threshold. Preliminary analyses indicated that there were no substantial systematic interactions in the accuracy or latency data as a function of the distance in number of asterisks from the response threshold, of whether the stimulus came from either the short or the long tail of the distribution relative to the response threshold, or of which response keys the 'low' and 'high' responses were paired with. On the basis of these analyses, and of the similar conclusions drawn by Ratcliff et al. (1999), the results that are plotted in the remainder of this section collapse across these variables for simplicity of presentation. The data presented in these plots has been smoothed with local regression to highlight the main trends in the data; the degree of smoothing was adjusted for each plot to achieve a compromise between minimizing non-systematic variability without introducing artifactual trends that were not present in the raw data. Note that, in contrast to the plotted data, several of these variables that were collapsed across were, however, included in the statistical analyses so as to quantify these effects and allow for the small but significant main effects caused by these variables to reduce the unexplained variance. These effects do not substantially alter the trends displayed in the figures.

The overall accuracy and latency for correct responses in the experiment were 84% (SE = 0.1%) and 600 ms (SE = 1 ms), respectively. These levels of performance indicate that despite the present task being easier in some respects than the task employed by Ratcliff et al. (2009)—for instance, because the 'low' and the 'high' conditions did not overlap and lead to incorrect responses that were optimal in a statistical sense—participants were still generally well below the performance levels reported in that study. This outcome is attributable primarily to the present analyses applying to the first set of trials designated as practice in the Ratcliff et al. experiment, during which participants become familiar with the task. To a lesser extent, this may also be due to the feedback presented at the end of the blocks, which encouraged participants to respond

more rapidly when their accuracy exceeded 90%; however, this instruction was only given at the end of 23% of the blocks because participants generally found it difficult to reach the 90% threshold with only 1,200 trials' worth of experience on the task.

Figure 8.1 and Figure 8.2 display the accuracy and correct and incorrect latency data as a function of the absolute distance in number of asterisks, from the response threshold, respectively. Both of these plots show that performance improved towards an asymptotic state as the distance from the response threshold increased. The latencies for correct and incorrect responses are quite similar up to a distance of approximately 12 asterisks from the response threshold, although there is a small latency increase for incorrect responses. Incorrect response latencies to numbers of asterisks that were more than 12 distant from the response threshold did not vary monotonically; however, limited weight should be placed on these data because they are based on only 240 observations in total. Figure 8.3 presents a combined latency-accuracy plot based on the data from Figures 8.1 and 8.2. Interestingly, and in contrast to the separate plots of accuracy and latency, this plot shows that latency differentially increased rather than asymptoted for correct responses as performance on the accuracy measure approached ceiling. Again, the data from the incorrect responses largely parallels that for the correct responses until the distance from the threshold becomes extreme and the remaining portion of the curve was not very reliable.

The skewness of the latency distributions was also examined to determine whether it conformed to standard patterns reported in other studies and if this measure, which has previously been argued to be more sensitive than analyses based on mean performance, provided additional insight into participants' behavior (Heathcote, Popiel, & Mewhort, 1991). The distribution of correct and incorrect latencies is skewed to the right, as illustrated in the density plots presented in Figure 8.4. An examination of skew as a function of distance from the threshold, which is presented in Figure 8.5, also shows that that skewness generally increased as a function of distance for correct responses,[4] except for very large distances for which the plotted curves are based on

[4]Except for distances that were very close to the threshold, there were insufficient numbers of incorrect responses to derive reliable estimates of the skewness of the error latencies.

very few data points (e.g., only 165 data points were included in the calculation of skew for correct responses for the plotted data corresponding to a distance of 26 asterisks from the response threshold).

In additional analyses not included here for the sake of brevity, it was also confirmed that the latency-accuracy and hazard functions for the present data set were generally similar to those reported by Ratcliff et al. (1999).

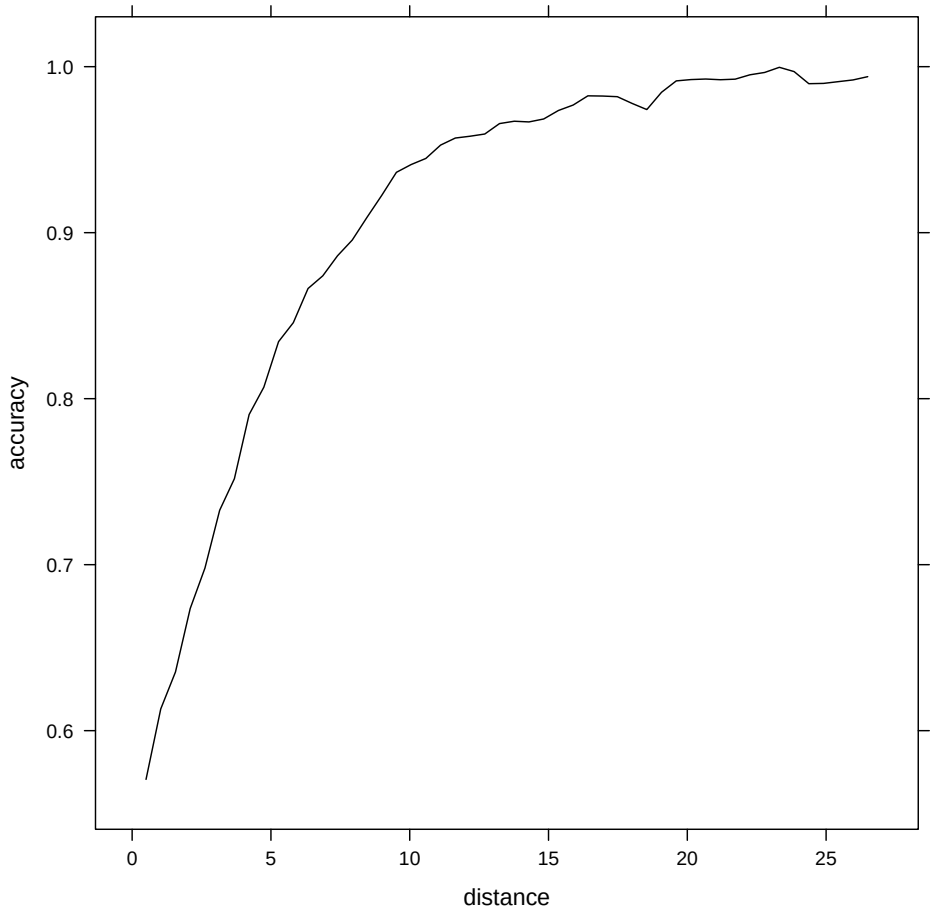


Figure 8.1: Plot of accuracy as a function of the distance in number of asterisks from the response threshold.

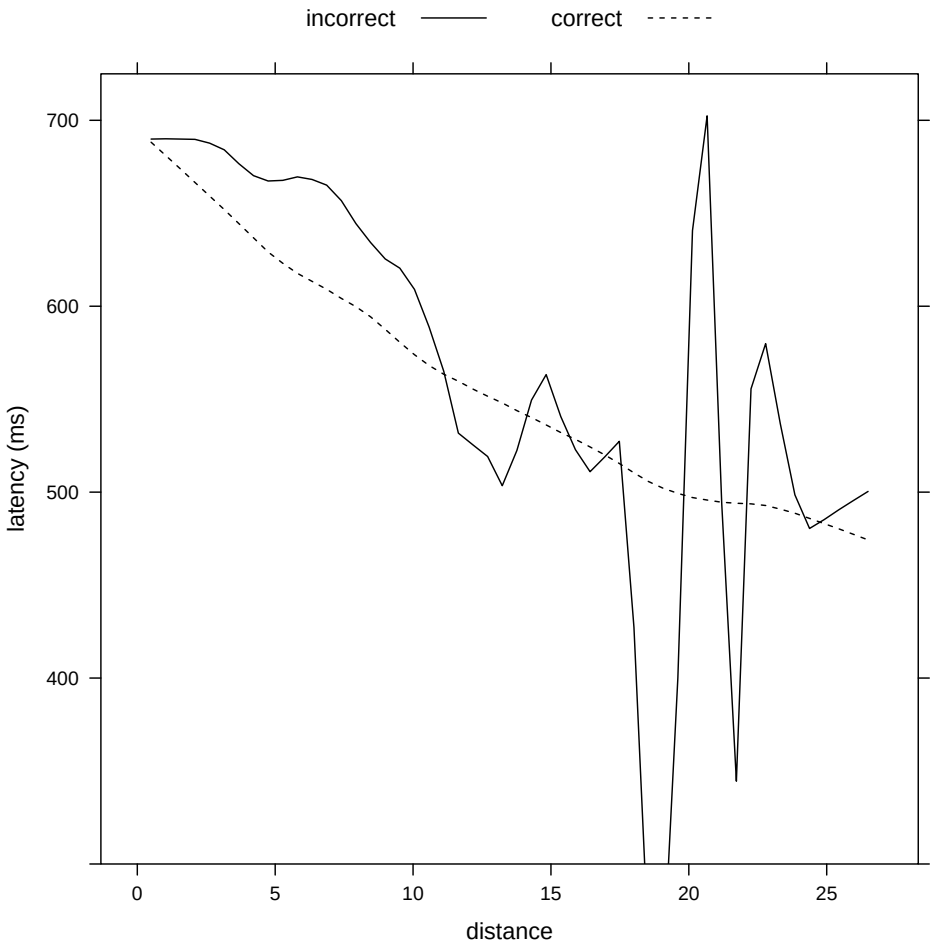


Figure 8.2: Plot of latency as a function of the distance in number of asterisks from the response threshold. Separate lines show the data for correct and incorrect responses.

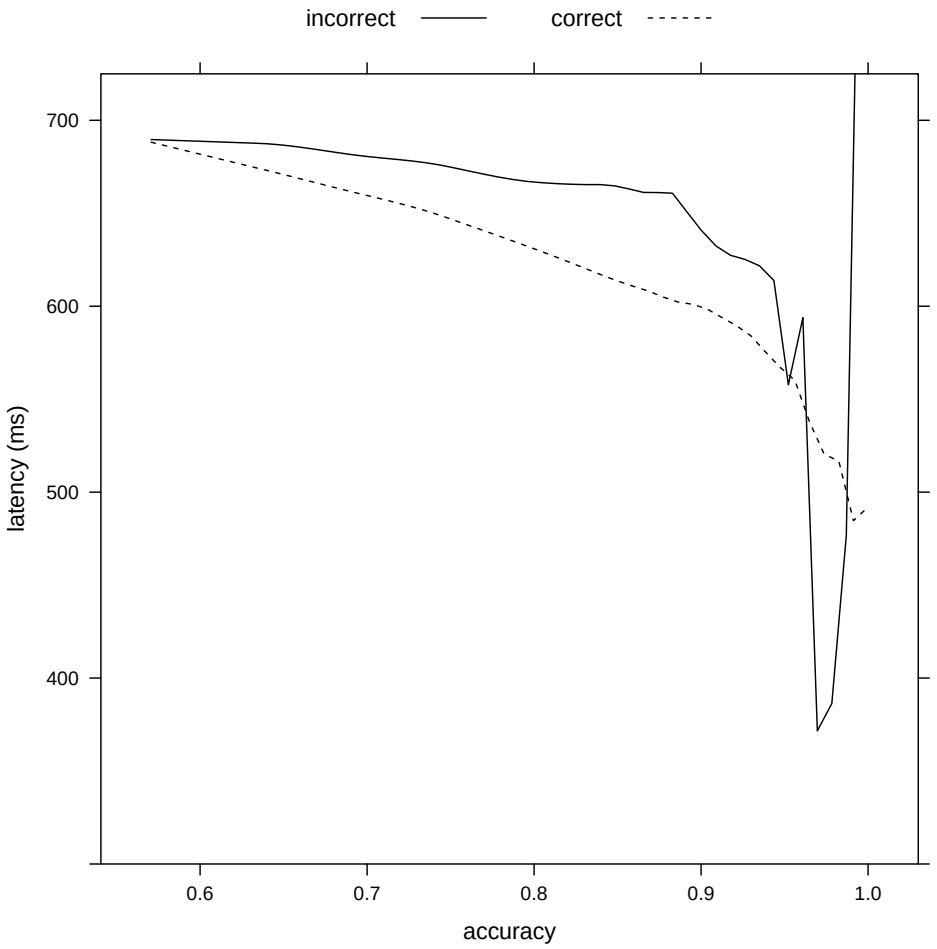


Figure 8.3: Plot of latency as a function of accuracy generated by combining the results from Figures 8.1 and 8.2. Separate lines show the data for correct and incorrect responses.

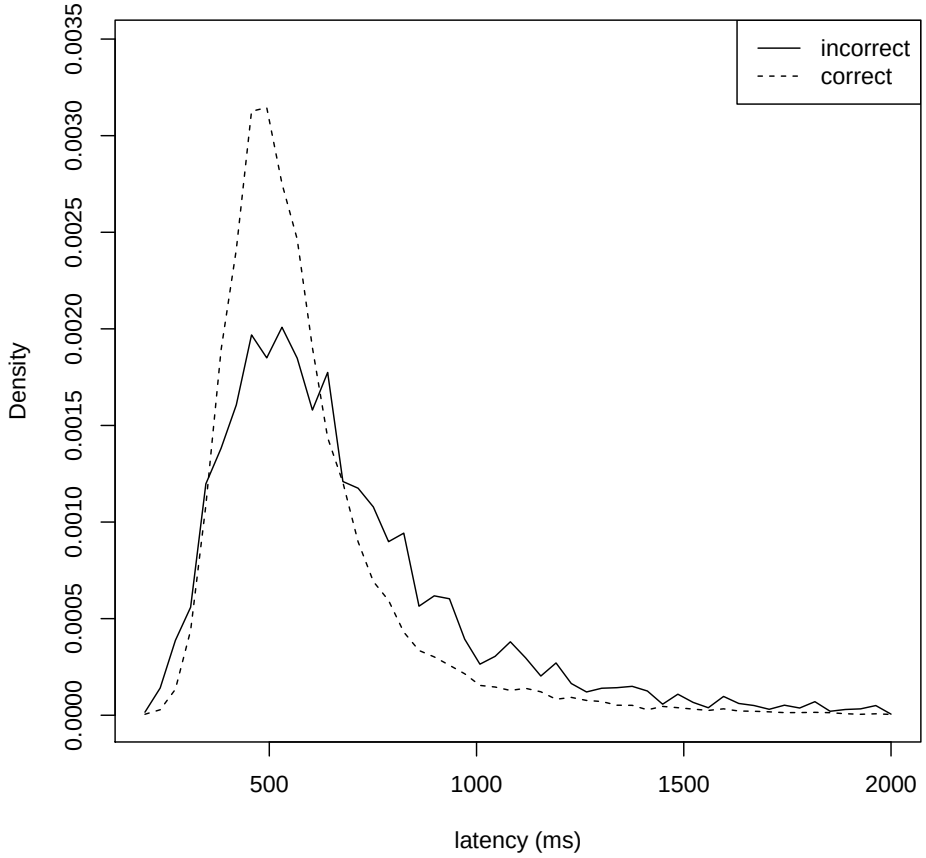


Figure 8.4: Distribution of latencies for correct and incorrect responses.

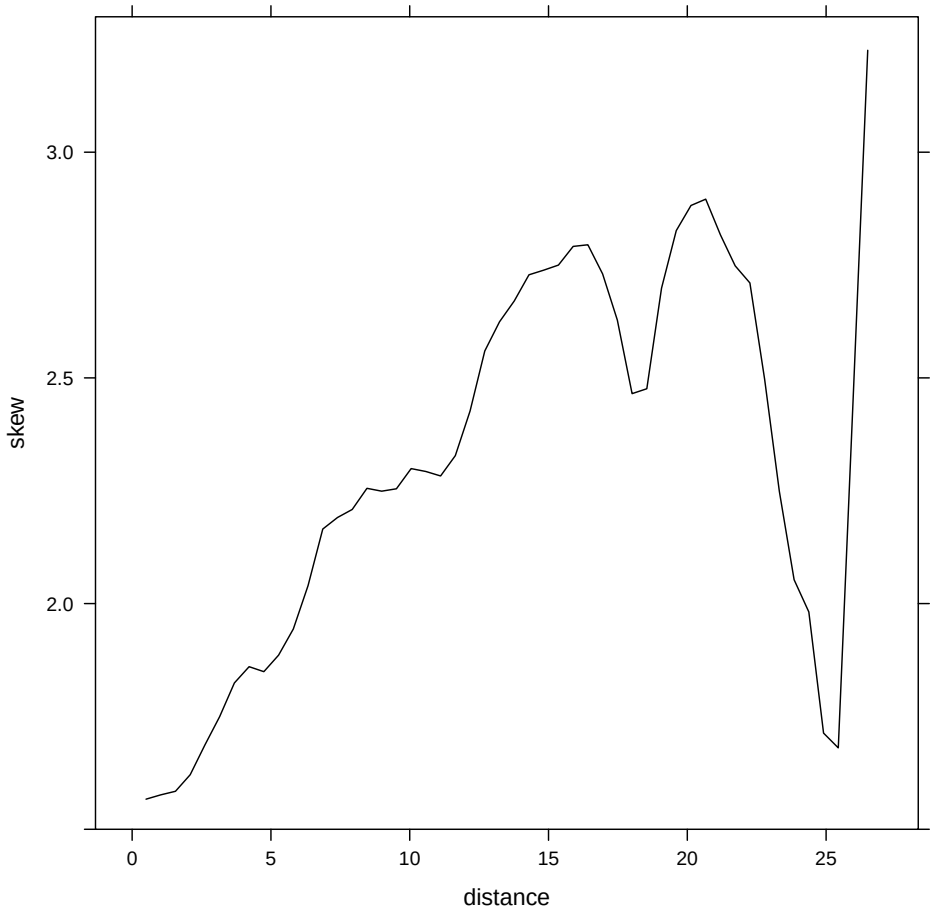


Figure 8.5: Degree of skew in the distribution of latencies for correct responses as a function of distance in number of asterisks from the threshold.

Sequential effects. The trial-by-trial sequential effects in the constant threshold conditions were examined in several ways to evaluate how performance changed as a function of previous experience. First, overall accuracy and latency were examined as a function of trial rank (i.e., number), and are plotted in Figure 8.6 and Figure 8.7; for the moment, the discussion focuses only on the constant threshold condition (same–same) and a detailed discussion of the other conditions is deferred to a later section. These data show that whereas participants exhibit near-asymptotic performance in terms of accuracy from the beginning of the experiment relative to their accuracy levels at the end of the experiment, the same was not true for the latency data. Initially, these data show an increase in latency, which occurs when the highest proportion of participants received feedback to "respond more acurately even if it means responding more slowly." As the accuracy level stabilizee after the first few blocks, substantial decreases in latency were observed throughout the remainder of the experiment. The direction and overall characteristics of these functions were also generally recapitulated in the data from the individual participants. Similar findings were also observed in the other conditions of the experiment until near the change in the threshold value. Note that in those other conditions, the effects of the change in threshold emerge slightly before the actual change in the graphs (trial rank = 400) because of the local regression smoothing used to highlight the main effects of interest. Note also that these sequential effects appear to be present across a range of initial speed-accuracy trade-offs and proficiency that resulted from testing different participants in the different conditions. However, the extent to which this should be viewed as a strong inferential claim is limited because the average latency for correct reponses in the first mega-block was not substantially different across conditions. This was determined by inspecting the box-and-whiskers plots in Figure 8.8 and noting that the median performance in each condition fell within the 25th and 75th percentiles of the latencies in the other conditions.

To gain further insight into how adaptation occurs in the system, additional analyses examined how performance changed as a function of number of blocks for stimuli of differing

distance from the threshold. To increase the reliability and simplicity of these plots, stimuli of similar distance from the threshold were binned together each time the distance from the threshold increased by 5 (i.e., items were binned into groups of 1–4 asterisks distant from the response threshold, 5–9 asterisks distant from the response threshold, etc.), and the data were averaged at the block as opposed to the trial level. Figures 8.9 and 8.10 plot the results for the accuracy and the latency data for correct trials, respectively. For both of these metrics, performance generally continues to improve throughout training, although to a lesser extent in the accuracy data, which also reach ceiling performance for many of the groups and did not show a substantial improvement for the group that was closest to the response threshold. The latency data also appear to show the greatest improvements as a function of practice for the groups which are furthest from the threshold. This fact, coupled with the near asymptotic accuracy levels for those groups, suggests that error feedback primarily serves to enhance the speed of responses that are already highly accurate in this task, and that the groups that are closest to the threshold are those that show the smallest increases in performance both with respect to accuracy and latency.

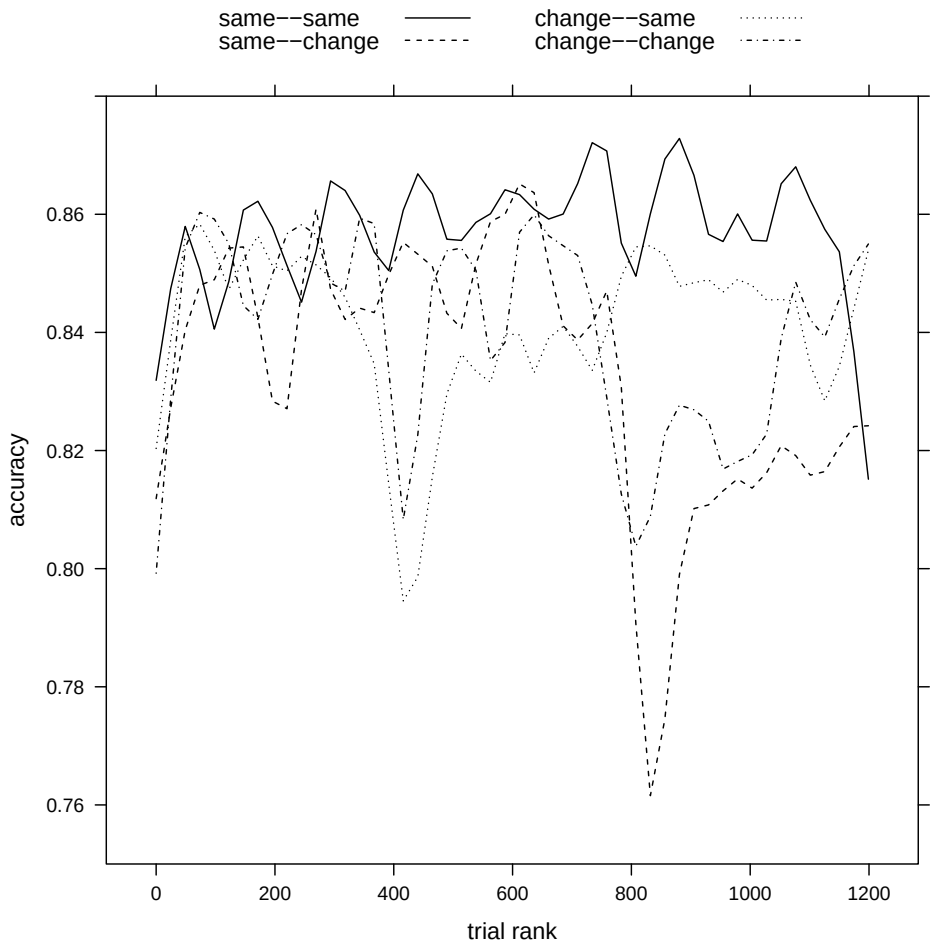


Figure 8.6: Accuracy as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks.

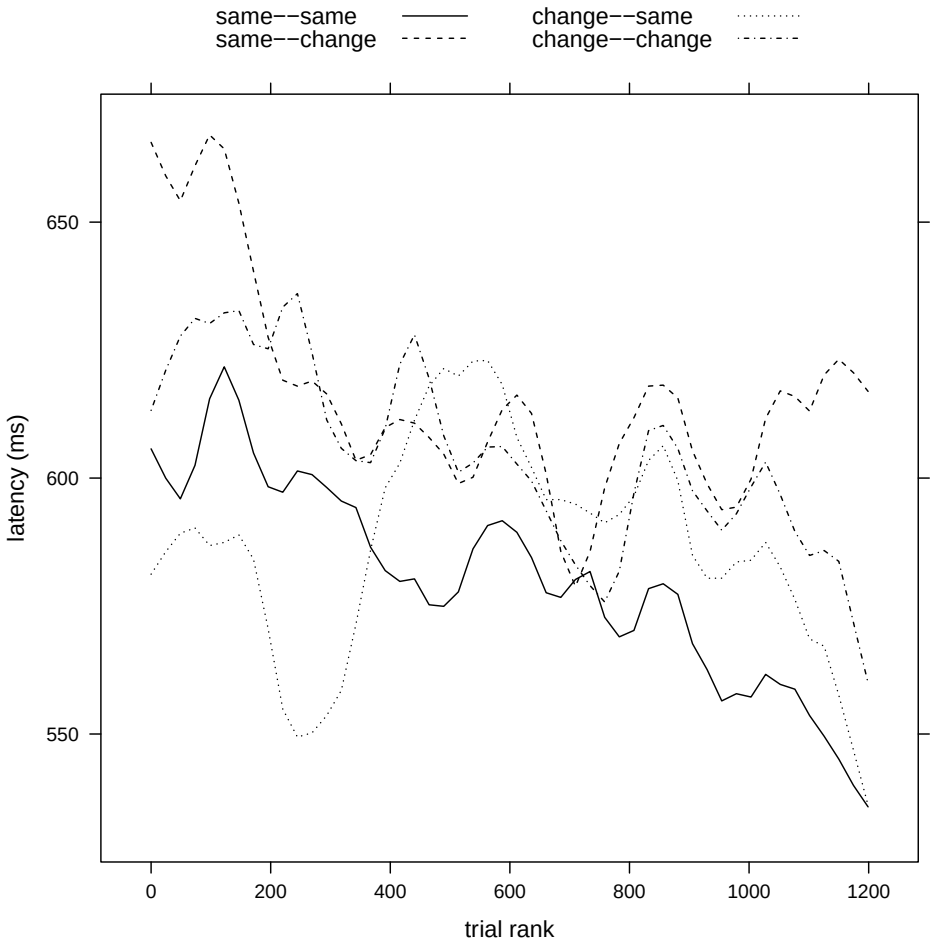


Figure 8.7: Latency for correct responses as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks.

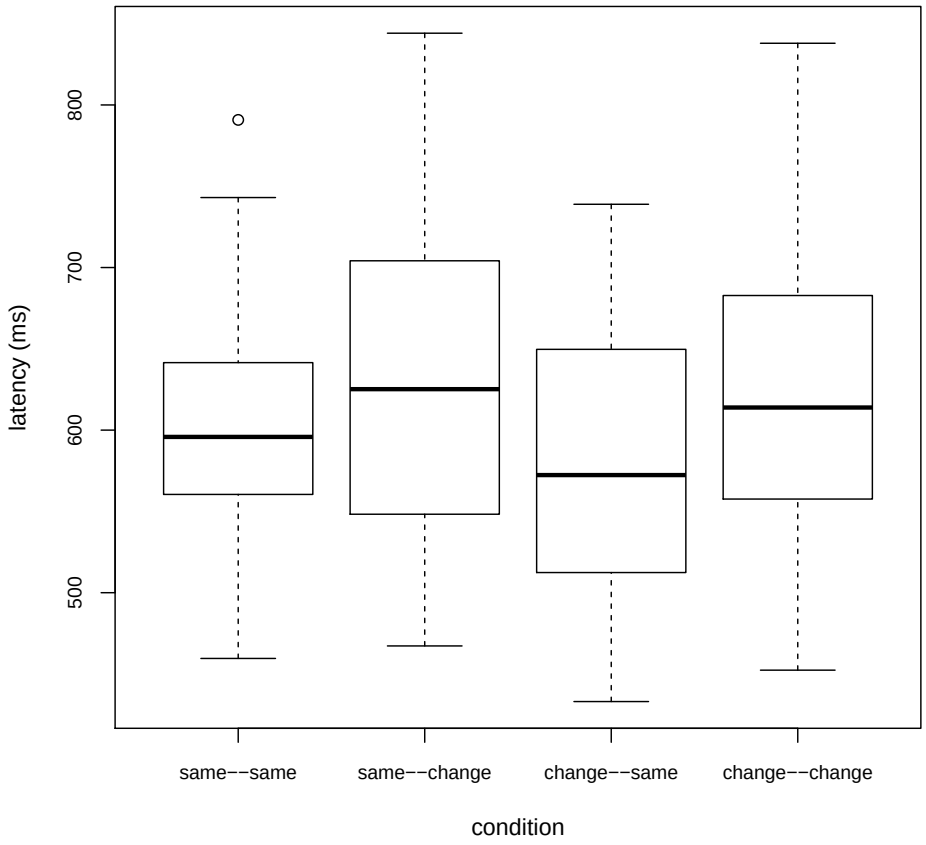


Figure 8.8: Boxplots of the latency data for correct responses for the first 400 trials in each condition.

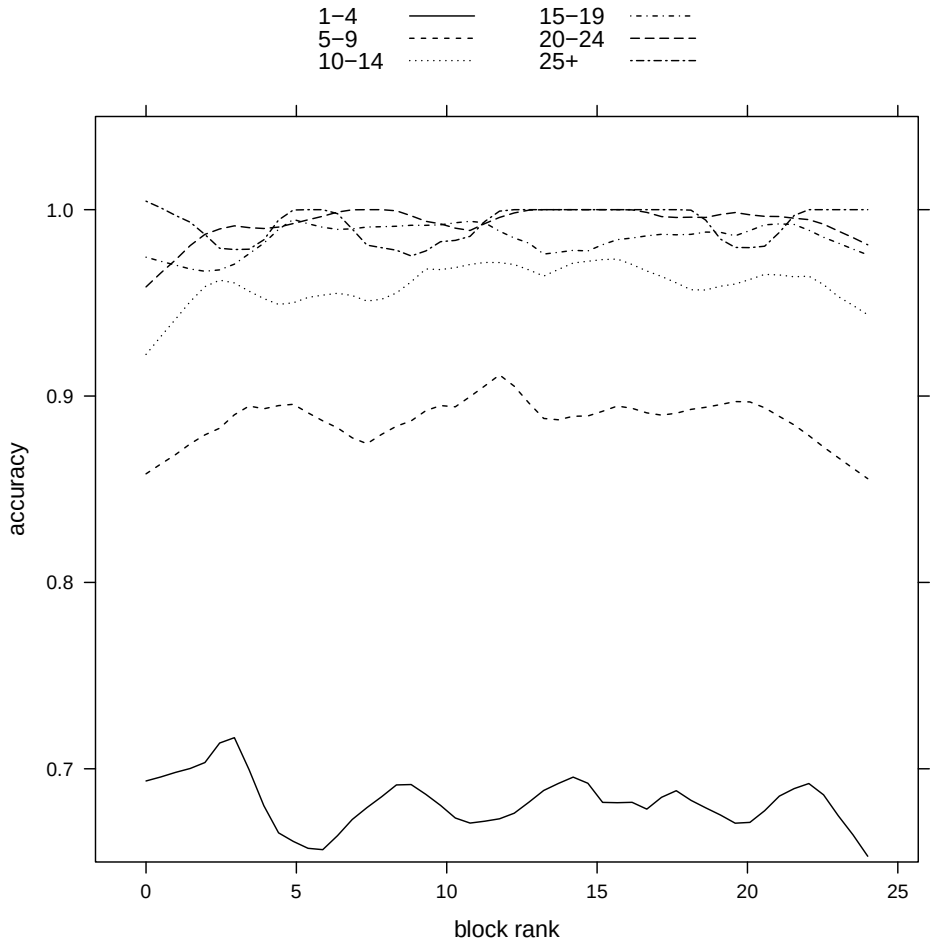


Figure 8.9: Accuracy as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines.

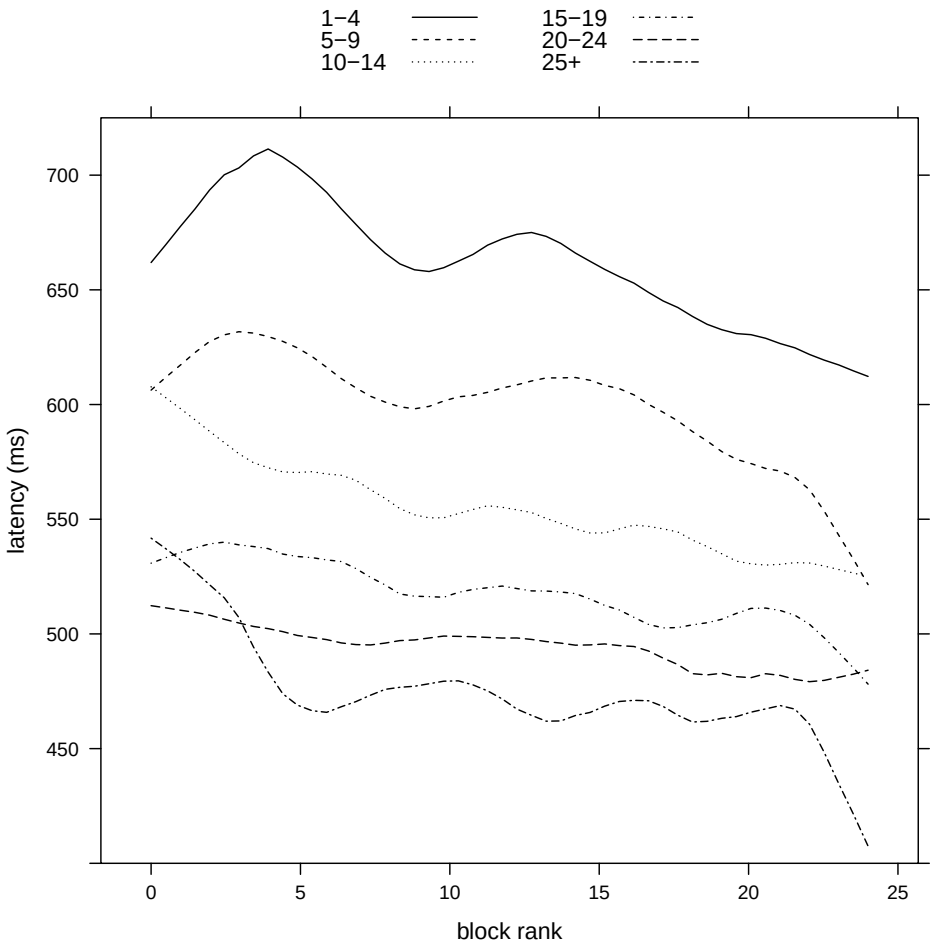


Figure 8.10: Latency for correct responses as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines.

Having established that participants do show adaptive effects throughout the experiment, further analyses evaluated which characteristics of the task environment influenced performance and which characteristics were predictive of trial-by-trial sequential effects. A mixed-effect regression model (Baayen et al., 2008) was used to quantify the effects of several factors on response accuracy and on the latency of correct responses. Several predictors related to general task environment and to the current trial were entered, including: the button used to signal a 'high' response (button), with the baseline set to 'high' responses being signaled by pressing '/' with the right hand; whether the stimulus came from below or above the threshold separating low or high responses (fstim), with the baseline set to the high responses (which were also more frequent in number); the distance in number of asterisks of the presented stimulus from the response threshold (distance); the duration of the fixation stimulus (fix); and the trial rank (trial). The influence of several characteristics of the participants' most recent and more extended experience with the task was assessed by including variables reflecting the participants' performance both on the preceding trials and during the previous block. Measures of performance on the previous two trials included the distance of the preceding stimuli from the response threshold (distance), the accuracy of the preceding response (acc), the latency of the preceding response (rt), and whether the preceding stimulus (repfstim) or response (represp) were repetitions of the stimulus and response presented on the current trial.[5] Subscripts on these variables indicated the lag, in number of trials, from the current trial. Measures of performance from the previous block included the accuracy (acc_b) and latency (rt_b) and whether the participants received feedback indicating that they should respond more accuractly (moreacc_b) relative to the baseline that they were told to respond more quickly. A general linear mixed model for binomial distributions and associated parametric tests were used to evaluate the effects of these predictors on accuracy. Similarly, a standard linear mixed model and Markov Chain Monte Carlo simulations were used to generate

[5]Including these factors also has the added benefit of minimizing auto-correlation effects that account for substantial variance—more, in fact, than many of the characteristics of the current trial—that can impede drawing accurate inferences from the model (Baayen & Milin, 2010).

the p-values [denoted as p(mcmc)] used to evaluate the effects of these predictors on latency. The sign of the coefficients has been standardized such that a positive slope in accuracy and a negative slope in latency indicates a performance improvement on the corresponding dependent measure.

Accuracy. The coefficients for the predictors of the accuracy data are presented in Table 8.1. These results recapitulate, quantify and refine the sequential effects presented in earlier figures: slightly more accurate responses are made for the more-frequently presented type of stimulus, greater distances from the response threshold are associated with increased accuracy, and accuracy increases as participants complete more trials. Additionally, these results show that correct responses on previous trials increase the likelihood of a correct response on the current trial, slower latencies on previous trials are associated with increased accuracy on subsequent trials, and repetitions of the same stimulus increase the accuracy of subsequent trials. The non-significant effects also indicate that difficulty, as assessed by the distance from the response threshold, of the previous trial is not a strong predictor of accuracy on a subsequent trial, and repetitions of the same response from a previous trial predict a decrease in accuracy (this is potentially understood as resulting from residual priming for generating the same response independent of the evidence that is presented on the current trial). Additionally, the results show that the feedback provided after each block, which on average indicated that the participant should respond more accurately even if it meant slowing down, showed only a numeric but statistically non-significant trend to increased accuracy on subsequent blocks. This suggests that participants were either not attending to the block-level feedback or were unable to increase their accuracy beyond the levels they had already reached even with this feedback. Essentially the same pattern of results was also observed in a similar analyses involving all of the participants but only the data from the first mega-block, prior to any potential shifts in the threshold. Additionally, a similar analysis that included an interaction term between the previous trial's accuracy and trial rank, and for which only the predictors from the immediately preceding trial were included to minimize

collinearity, identified a non-significant trend towards a reduced influence of the accuracy of the previous trial as a function of increased experience with the task, $\beta = -0.0002, \mathrm{SE} = 0.0001$, z = -1.4, p = 0.16.

Table 8.1: Statistics for the predictors in the accuracy analysis

	β	SE	z	p	
(Intercept)	−0.7708	0.3513	−2.2	0.028	*
button	0.0826	0.0389	2.1	0.034	*
fstim	−0.3450	0.0376	−9.2	0.000	***
distance	0.2324	0.0044	52.9	0.000	****
fix	−0.0001	0.0003	−0.4	0.657	
trial	0.0001	0.0001	1.5	0.146	+
distance_1	−0.0027	0.0028	−1.0	0.328	
acc_1	0.0975	0.0507	1.9	0.055	+
rt_1	0.0002	0.0001	2.8	0.004	**
repfstim_1	0.2706	0.0380	7.1	0.000	***
represp_1	−0.3186	0.0370	−8.6	0.000	***
distance_2	−0.0043	0.0028	−1.5	0.125	+
acc_2	0.1205	0.0506	2.4	0.017	*
rt_2	0.0001	0.0001	1.5	0.145	+
repfstim_2	0.1148	0.0378	3.0	0.002	**
represp_2	−0.1657	0.0366	−4.5	0.000	***
acc_b	0.0102	0.0030	3.4	0.001	***
rt_b	0.0001	0.0002	0.5	0.606	
moreacc_b	0.0632	0.0492	1.3	0.199	

Note. + = p < .15, * = p < .05, ** = p < .01, *** = p < .001.

Latency for correct responses. The coefficients for the predictors of the latency data for correct responses are presented in Table 8.2. These results largely parallel those observed in the accuracy analyses with the following deviations: increased difficulty (in terms of reduced distance from the threshold) of the previous response predicted slower responses on subsequent trials, the facilitation associated with repetitions of the same type of stimulus were less robust, and repetitions of previous responses and the accuracy of previous responses reduced latency, but only to a significant extent for the predictors from the immediately preceding trial. This suggests that the decreased accuracy associated with the response repetition predictor in the accuracy analysis is at least partially due to a speed-accuracy trade-off. The performance from the previous block was more strongly predictive in the latency analyses, which showed that participants adjusted their average latencies on the basis of the feedback that they received. Again, the same pattern of results was also observed in a similar analyses involving all of the participants but only the data from the first mega-block, prior to any potential shifts in the threshold. However, this analysis did detect stronger effects of stimlus type repetition and of the difficulty of the preceding trials. Additionally, a similar analysis that included an interaction term between the previous trial's accuracy and trial rank, and for which only the predictors from the immediately preceding trial were included to minimize collinearity, found a significant effect of reduced influence of the accuracy of the previous trial as a function of increased experience with the task, $\beta = 0.05, \mathrm{SE} = 0.01$, t = -5.1, p(mcmc) < 0.001.

Table 8.2: Statistics for the predictors in the analysis of the latencies for correct trials

	β	SE	t	p(mcmc)	
(Intercept)	413.27	26.80	15.4	0.000	***
button	15.32	12.76	1.2	0.228	
fstim	27.87	2.75	10.1	0.000	***
distance	−8.16	0.18	−45.5	0.000	***
fix	−0.10	0.02	−5.3	0.000	***
trial	−0.03	0.00	−7.5	0.000	***
distance_1	1.03	0.19	5.5	0.000	***
acc_1	−14.22	3.45	−4.1	0.000	***
rt_1	0.10	0.00	23.4	0.000	***
repfstim_1	−3.71	3.26	−1.1	0.264	
represp_1	−12.27	3.27	−3.8	0.000	***
distance_2	0.19	0.19	1.0	0.292	
acc_2	−4.36	3.44	−1.3	0.203	
rt_2	0.08	0.00	17.5	0.000	***
repfstim_2	−10.01	3.26	−3.1	0.001	**
represp_2	−0.62	3.27	−0.2	0.874	
acc_b	0.47	0.22	2.1	0.034	*
rt_b	0.26	0.01	20.2	0.000	***
moreacc_b	38.02	3.37	11.3	0.000	***

Note. + = p < .15, * = p < .05, ** = p < .01, *** = p < .001.

Adaptive effects. The following section reports the adaptive effects that resulted from a change in the threshold across two adjacent mega-blocks.

To determine the degree to which participants were explicitly aware of the changes in feedback, a d-prime measure (Macmillan & Creelman, 2005) was derived on the basis of the number of participants who reported that they believed the threshold had changed throughout the experiment in both the constant threshold and variable threshold conditions. The resulting d-prime value was 0.4, indicating that participants were not, on average, aware of the change in feedback.

The overall effect of practice for the constant threshold condition and each of the variable threshold conditions is plotted in Figure 8.6 and Figure 8.7 for accuracy and latency for correct trials, respectively. These figures show that, despite the considerable variability within the conditions, there are clear, strong effects of threshold changes on both accuracy and latency, with immediate accuracy decreases and latency increases following a threshold change. Additionally, these results indicate that as participants gain more experience with the task, the degree to which they are able to adapt to threshold changes critically depends on their experience with previous threshold changes: when participants have previously encountered a threshold change (change–change) their adaptation following a second threshold change occurs at approximately the same rate as their initial adaptation to a threshold change. In contrast, participants who had completed two blocks with the same threshold and then encountered a threshold change (same–change) showed much more extensive degradation in their performance and were unable to return to a level of proficiency similar to that which they displayed prior to the threshold change. To evaluate whether different portions of the distribution were more affected by the change in threshold than others, the performance in each mega-block for each condition was plotted relative to a standardized response threshold. In these plots, all of the data were transformed to appear as if the initial threshold had been low for all conditions, the first threshold change had been to a high value, and if applicable, the second threshold change involved a switch back to the initial low threshold. These plots are presented in Figure 8.11 and Figure 8.12 for accuracy and correct

latency, respectively. These plots show that although only the correct responses for the central portion of the asterisk distributions were affected by the change in threshold, performance across the entire range of values was altered following a threshold change; albeit, to a lesser degree for more extreme values. The plots also show decreased overall performance and slower adaptation for the condition in which the threshold change only occured after two blocks at the same threshold (same–change).

Figures 8.13 and 8.14 plot accuracy and correct latency as a function of trial rank for different standardized numbers of asterisks and combinations of constant or changing response thresholds across mega-blocks to provide more detailed insights into the adaptive effects. The data from three groups of asterisks are presented: one group fell just below the initial low threshold (37–45), one group fell just above the initial low threshold and just below a later high threshold in conditions in which a threshold change occured later (46–54), and one group was well above the initial low threshold but was immediately above a high threshold if the threshold value changed (55–63). Data from the most extreme standardized numbers of asterisks were excluded because the rarity of these trials for any given trial rank was such that reliable curves could not be achieved without very high amounts of local regression smoothing. The results show that in both the accuracy and the latency data the adaptation that followed a threshold change occured over an extended number of trials. Specifically, performance generally did not approach an asymptotic level until approximately 100 trials after the threshold change. The magnitude and consistency of these changes was also larger for the two groups that always received consistent feedback and differed only in their distance from the threshold (37–45 and 55–63).

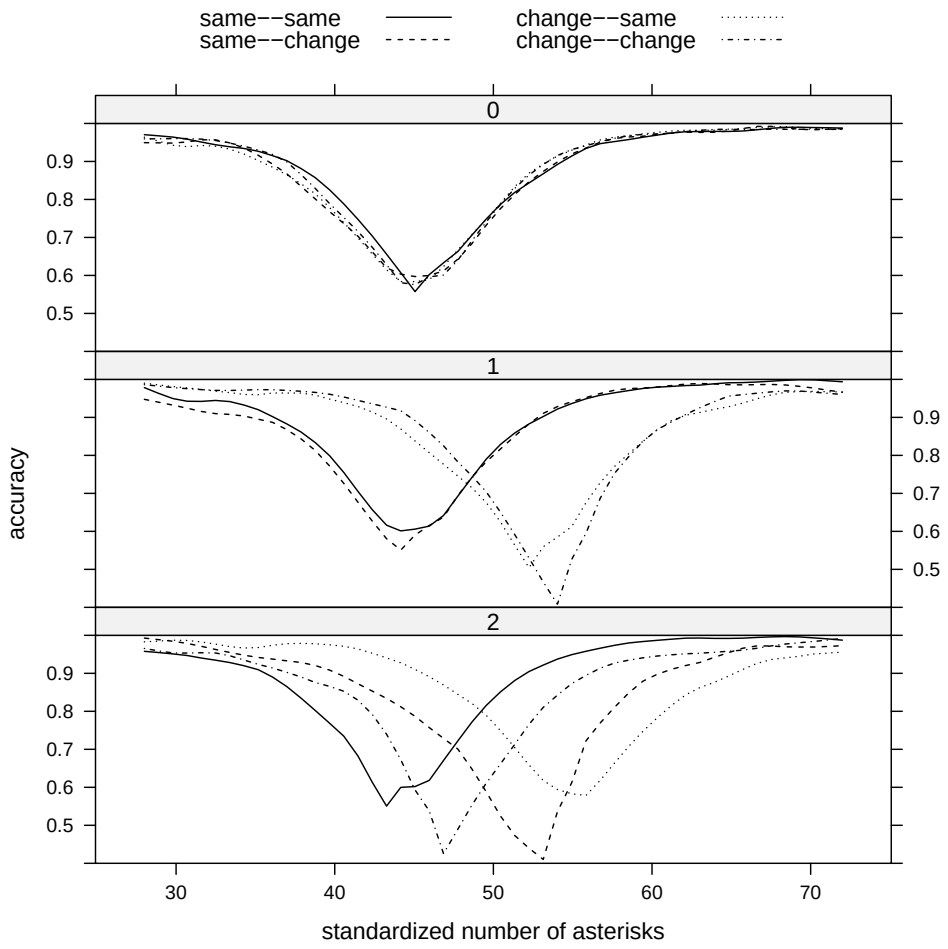


Figure 8.11: Accuracy as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions.

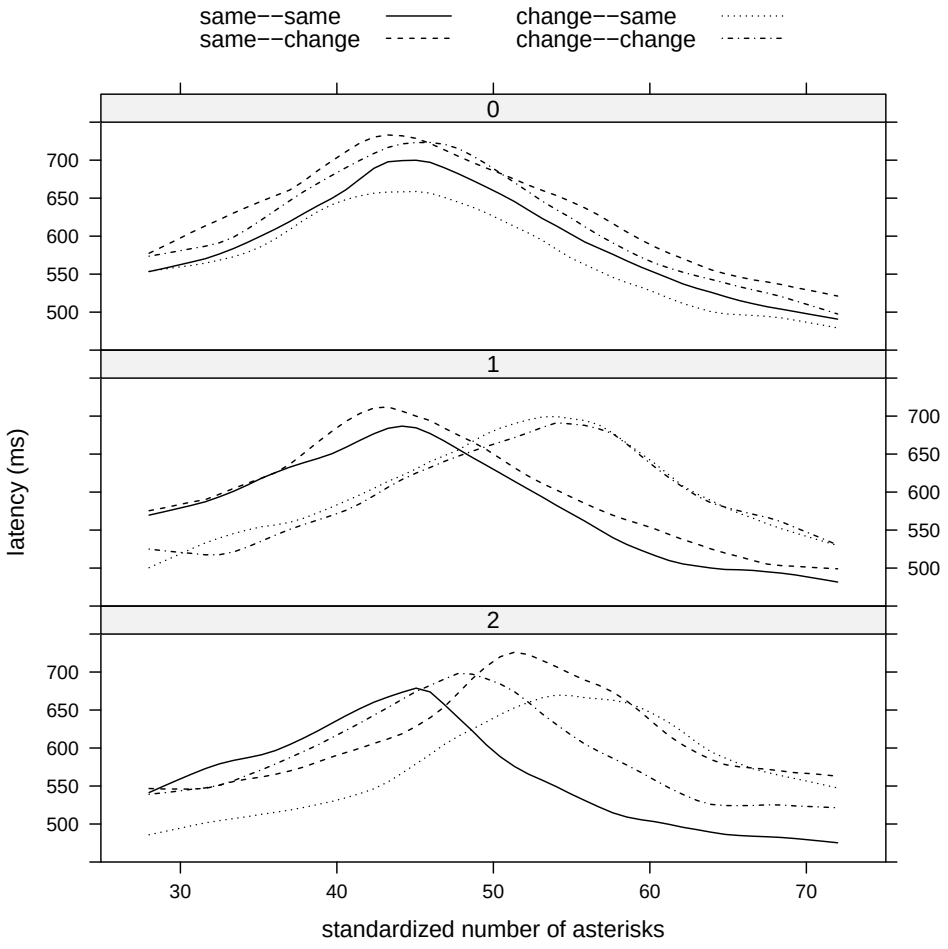


Figure 8.12: Latency for correct responses as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions.

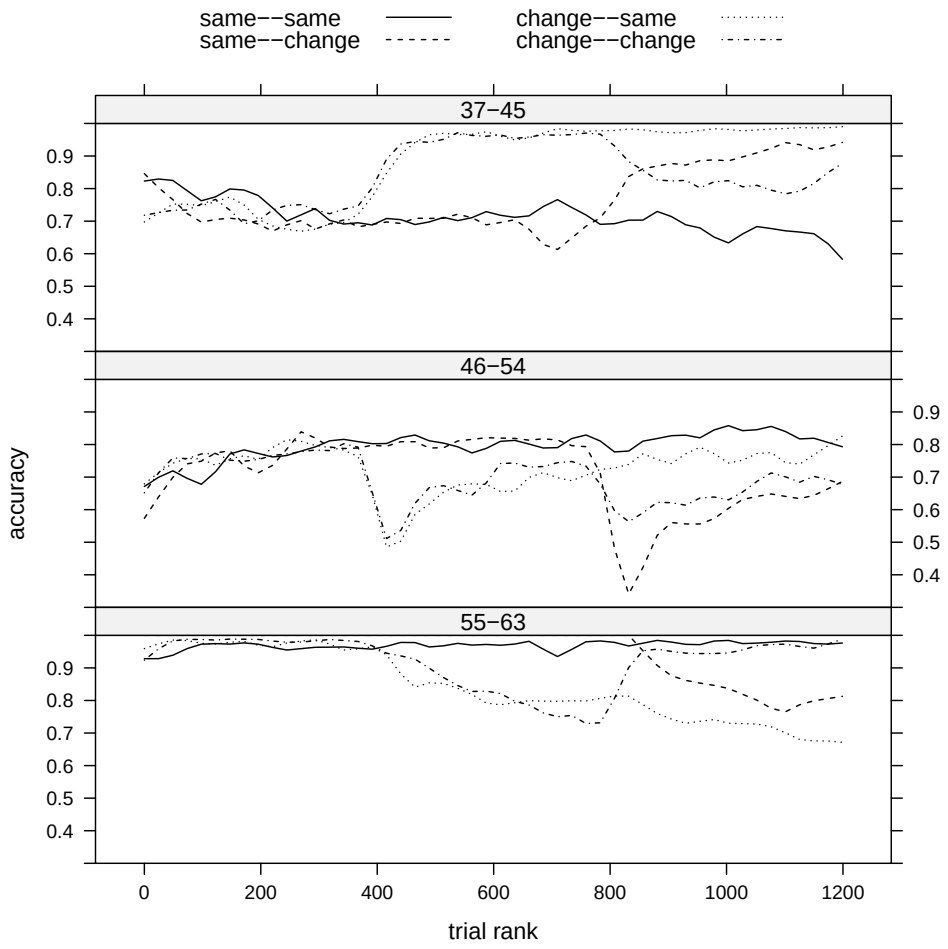


Figure 8.13: Accuracy as a function of trial rank. Separate panels present different groupings of standardized number of asterisks and separate lines within a plot represent the different conditions.

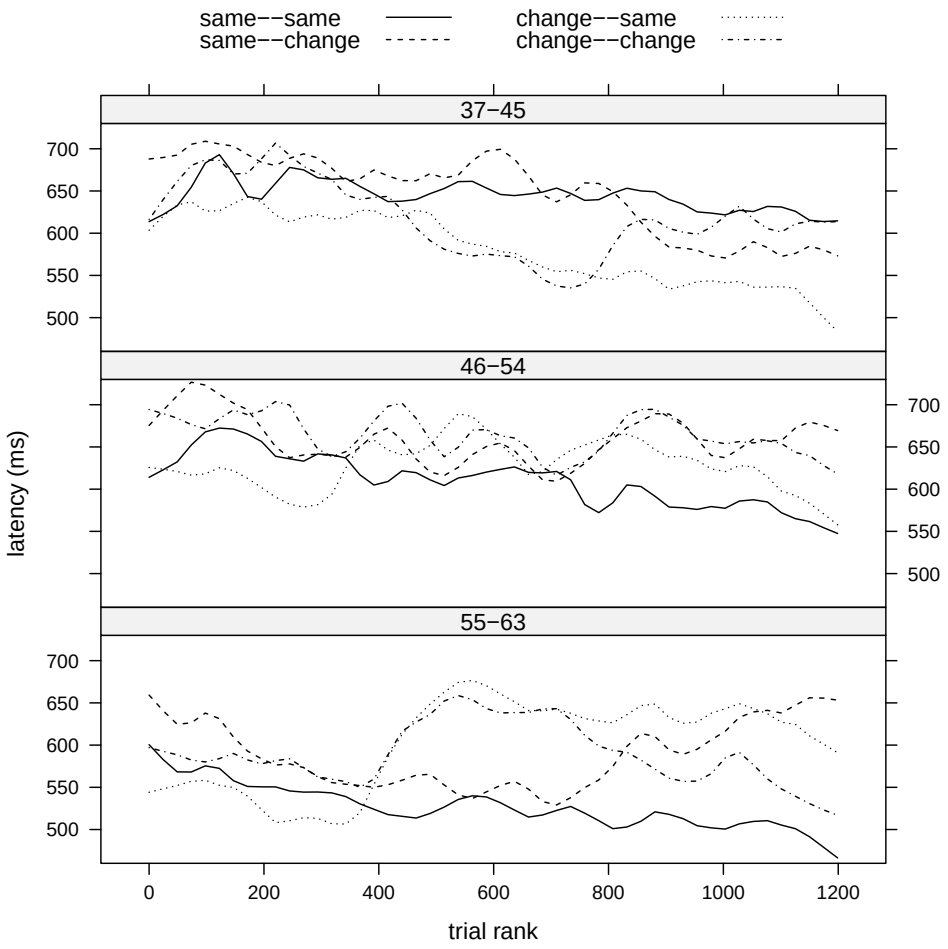


Figure 8.14: Latency for correct trials as a function of trial rank. Separate panels present different groupings of standardized number of asterisks and separate lines within a plot represent the different conditions.

Discussion

Despite many methodological similarities, the results of the present experiment diverge in several important respects from the results and associated modeling assumptions reported by Ratcliff et al. (1999). First, with respect to initial proficiency, participants were able to pre-configure their response selection system on the basis of the instructions alone so as to make accurate responses at above chance levels from the very first practice trial. Additionally, the bulk of their improvements in accuracy occured during the practice block, with accuracy increasing by approximately 20% following between the first practice trial and the first experimental trial (i.e., over the first 10 trials of the experiment). Nevertheless, participants continued to show small overall improvements in accuracy and substantial improvements in latency as they completed more trials. Together, these results suggest that some middle-ground between a system that is not at all set up to complete the task from the outset and a system that has already reached an asymptotic level of performance should be the target for connectionist simulations, and not either of those extremes as examined in the connectionist simulations reported by Ratcliff and colleagues. Moreover, it suggests that such a system should be biased towards making correct responses that are slow rather than making incorrect responses that are fast. As a result, the trial-by-trial effects should be more readily detected in the latency data, both because this is the metric that can show the greatest improvement and because of the superior statistical power associated with continuous as opposed to binomial variables. Indeed, this was essentially what was observed in the present experiment. These data also highlight the valuable information that can be gleaned by considering early performance and not simply discarding it, as was done in the original study.

In terms of overall performance, participants showed similar qualitative effects on many of the metrics that were examined in both studies, including how accuracy and latency varied as a function of distance from the threshold and in terms of the skewness of the latency distribution for correct and incorrect trials. The present results also provide additional details to those

provided in the original report: for instance, by showing that skewness increased as a distance from the threshold increased and response latencies decreased, as has been reported across a wide range of cognitive tasks (Hockley, 1984; for additional discussion see Mewhort, 2005), and by quantifying the effects of various measures related to the previous trial on performance on subsequent trials. In particular, the present study showed that repetitions of a stimulus lead to more accurate responses for that stimulus on a subsequent trial (suggesting priming or other facilitation for the activation of the representation associated with that stimulus and/or response; Plaut & Booth, 2000), whereas repetitions of a response led to less accurate but faster responses on subsequent trials (suggesting that residual activity of the representation of the previous response influenced performance on the following trial). It also showed that responses were more accurate if the previous response was slow and responses were faster if the previous response was accurate, which indicates the presence of a speed-accuracy trade-off at the trial level. Moreover, the present study identified several small effects associated with the second most recent trial. These effects (or, conversely, the absence of strong effects) for less recent trials can provide additional fine-grained insights into how the response selection system adapts on the basis of its prior experience.

Additionally, there are a number of discrepancies in the sequential effects that were reported by Ratcliff et al. (1999) and those that were observed in the present experiment, which are especially informative given the motivations for this study. Of particular noteworthiness is the observation that participants showed significant effects of the accuracy of the previous trial in altering both the accuracy and the latency of a subsequent trial. This is exactly what would be predicted by a model that employs an error-driven learning algorithm. These results also parallel those that have been observed in other experiments, such as lexical decision (Armstrong & Plaut, 2011; Armstrong, Tokowicz, & Plaut, 2012), and which have been argued to operate even in improving performance using internal detectors of potential errors in the absence of accuracy feedback from the environment (Joordens et al., 2009; for supporting empirical evidence, see the

control conditions from Armstrong et al., 2009).

Several factors likely contribute to this discrepancy. One of these factors is that the present work employed more powerful statistical methods from those employed in the Ratcliff et al. (1999) study. Given the relatively small—but statistically significant—magnitudes of many of the effects that were observed in the analysis, such methods may be necessary to detect the hypothesized effects of error-driven learning on a trial-by-trial basis. Another factor is that the present experiment focused on early learning before performance reached an asymptotic level. During this period, error-driven learning is hypothesized to have a larger impact on overall performance because the number and presumed magnitude of participants' errors (in terms of how close they were to making the correct response as opposed to the incorrect response) is considerably higher than later in the experiment. This hypothesis is supported by the interactions between the accuracy of the previous trial and trial rank in predicting the accuracy and latency for a subsequent trial. For both accuracy and latency, a numeric trend towards a decreased effect of the previous trial's accuracy in predicting subsequent performance was observed; albeit, this trend reached statistical significance only in the latency analysis. These results suggest that the influence of the previous trial's accuracy is reduced with practice and also likely contributes to the discrepancy between the results from the Ratcliff et al. experiment and from the present experiment. Additional factors, such as the inconsistent nature of the feedback provided in the Ratcliff et al. experiment (e.g., the same number of asterisks could have been generated by sampling from either the high or the low distribution and so different feedback signals can be associated with the same stimulus) may also have facilitated the detection of these effects in the present study if participants adjust the magnitude of their error corrections on the basis of the reliability and consistency of the feedback used to make prior corrections (Hinton, 1989; Jacobs, 1988).

The adaptive effects observed in the present work and in the Ratcliff et al. (1999) study also differed in several respects. Most importantly, the rate of adaptation to an unexpected change in the threshold was relatively slow and approximately 100 additional trials were necessary to reach

a new asymptotic level of performance in the present experiment. This contrasts with the results reported by Ratcliff and colleagues, which showed that participants adapted to a change in the threshold after approximately 5-15 trials, depending on the distance of the presented number of asterisks from the threshold. This discrepancy most likely results from the fact that participants in the Ratcliff et al. study were given explicit knowledge about the details of the task and had extensive experience with the distribution switches and when they occured. Consequently, they developed additional explicit strategies to adapt to these changes beyond attending to the simple statistical structure of the stimuli and associated feedback, which was made more likely by the methods employed in the present experiment. This claim is supported by the fact that in the present experiment, participants who encountered two changes in the threshold as opposed to one showed fairly comparable patterns of adaptation after each switch as opposed to a substantially facilitated adaptation (although admittedly, participants in the Ratcliff et al. study saw 32 distribution switches, which is considerably more than were given in the present experiment). Moreover, the results showed that a later switch in the threshold is associated with substantially greater performance impairments than an earlier switch. These effects are consistent with the notion that participants' response systems were becoming 'entrenched' and optimized for responding to one type of stimuli at the expense of being able to adapt to a sudden threshold change (Munakata & McClelland, 2003). If similar adaptation mechanisms were involved in optimizing performance in both tasks, these data would have suggested that the Ratcliff et al. experiment, which involved the presentation of 1,200 trials at a constant threshold before any threshold changes, should have produced adaptive effects that were slower (or at least of comparable speed) to those observed in the present experiment—not an order of magnitude faster.

In addition to these critical discrepancies, it is also worth noting that the present work also found that the reliability of the error data was relatively low except for near the response threshold. This suggests that these data should be underweighted when comparing them to model data relative to the more reliable data from correct responses. Such a position does, however, contrast

with the claims of Ratcliff et al. (1999), who viewed these data as a strong test of a model's similarity to the behavioral data. Given that these data are not reliable and replicable across similar tasks, this does not appear to be a reasonable golden standard for model comparison in a general sense, at least without a better understanding of why these patterns differ.

Taken together, the numerous discrepancies between the Ratcliff et al. (1999) study and the results observed here and in other related work (Armstrong et al., 2009) clearly question the representativeness of the Ratcliff et al. results as a gold standard for model comparison. This in turn motivates the re-evaluation of the similarity between a connectionist model's simulated performance and these behavioral results. Such a comparison is the subject of the next chapter.

Before turning to the simulations, however, it is worth noting that the present results should not simply be viewed as a replacement for the original gold standard set by Ratcliff et al.'s (1999) original experiment. Ultimately, a general model of response selection—connectionist or otherwise—should be capable of accounting for those data as well, although doing so may involve incorporating representations and processes of explicit strategic adaptation. Furthermore, as was articulated at the beginning of the chapter, the current state of affairs in the response selection literature has been to focus on modeling the details of very simple tasks. This has, however, been at the expense of developing theories that can account for performance in tasks that may involve more complex representations both of the underlying evidence that drives responses and of the responses themselves (e.g., lexical decision, semantic categorization, word and nonword naming). The new data described in this chapter should therefore not be viewed as 'the' gold standard for model comparison, but as one of many such sets that should be applied to test both the generality and precision of a given model. Nevertheless, as a starting point, the present data and task do have several strengths for studying particular aspects of the response system. This is because, as noted by Ratcliff and colleagues (but apparently not realized in their own study) they allow for the use of very simple representations of stimuli and responses, and because they maximize the likelihood that participants will employ simple and domain-general trial-by-trial

learning mechanisms. Consequently, the main challenge is to develop a system that can capture, explain, and predict effects related to these data that can also be extended to other ‘gold standard’ data sets from a range of tasks that better tap other aspects of response selection.

Chapter 9

Evaluating an Improved Model of Response Selection: Connectionist Simulations

The present chapter aims to re-evaluate how well a model of response selection based on a more biologically plausible connectionist formalism fares at capturing the behavioral effects reported in the previous chapter. The results show that the model does suprisingly well and captures many of the sequential and trial-by-trial effects from the behavioral data that are outside the scope of standard models of response selection, although there are a few inconsistencies in the details. Possible explanations and future studies investigating these issues are discussed.

9.1 Developing an Improved Connectionist Formalism

The behavioral results reported in the previous chapter call into question the original criticisms raised by Ratcliff et al. (1999) regarding the suitability of the connectionist framework as a general model of response selection. Consequently, whether or not a connectionist model can successfully model response selection is an open empirical question to which this chapter aims to

provide an initial answer. To this end, it is worth considering the exact connectionist formalism that should be employed for this purpose in light of the broader goals of the modeling enterprise beyond response selection.

One option would be to employ the standard connectionist formalism as outlined by Rumelhart and McClelland (1986a). Indeed, an initial attempt to model response selection using a model grounded in this formalism has shown considerable promise at capturing many of the types of effects that were observed in the empirical work, and served as an important foundation for this work (Mickey, personal communication, May 5th, 2011). Much of the work that has been reported to date in the dissertation has, however, employed a modified version of this standard formalism based on additional neurobiological constraints. For the sake of homogeneity and consistency with this work and the underlying neurobiology, the present work therefore uses a simliar formalism to that from Part I.

The use of this more biologically-plausible formalism also has the advantage of gaining many of the properties of the leaky accumulator models, either because they are specified explicitly or because they appear to be emergent characteristics of the formalism. For example, between-layer connections are positive in the leaky accumulator framework; the same is true in the more biologically-plausible formalism. Likewise, the units in the settling dynamics simulation that represented inconsistent semantic features did not develop substantial positive connections between one another and were able to inhibit one another indirectly via the inhibitory unit; this is similar to the restriction in the leaky accumulator model that required negative weights between inconsistent responses. However, the former model was more general in that it has already been applied in simulations employing distributed representations (particularly for the semantic 'outputs'). Two additional parameters in the leaky accumulator model that relate to the leakage of a unit's activation and to the reduction in noise over time are primarily concerned with modeling the asymptotic activation levels for the response units and the detailed characteristics of the latency distributions (Usher & McClelland, 2001). For simplicity, these parameters are set aside

for the moment because they are likely to play a reduced role in influencing the model's overall performance in simulating a speeded task in which the model is trained to respond as quickly and as accurately as possible.

There is, however, one important modification in the leaky accumulator model that was clearly not present (either because of an explicit restriction or as an emergent characteristic) in the earlier simulations: the use of a threshold-linear unit activation function (i.e., a function in which no output is produced until the net input reaches a specific threshold, after which point the output is a linear function of the net input). Rather, the earlier work employed a standard sigmoidal activation function. Additional research, however, suggests that neuron activation functions more closely resemble the threshold-linear activation function used in the leaky accumulator model and its variants (Usher & McClelland, 2001; Bogacz et al., 2007; Purcell et al., 2010). Indeed, the use of a sigmoidal function, which was motivated in part by sigmoid-like neural activation dynamics when neurons were injected with high levels of (input) current in vitro (Rumelhart & McClelland, 1986a), appears to tap an upper-bound on a neuron's activation dynamics that is not typically reached in vivo. Consequently, a threshold-linear function may be a useful addition to the more biologically-plausible variation of the connectionist framework that was introduced in the earlier sections, despite the computational advantages of sigmoidal units (e.g., they have inherently stable states at both low and high activation levels). Furthermore, making this modification would increase the similarity between the standard connectionist framework and the leaky accumulator framework, which has shown considerable success at fitting some of the detailed empirical effects associated with response selection.

Before adopting this change, it is worth briefly discussing two issues with this proposal. First, as noted earlier, the use of a threshold-linear activation function would not satisfy the requirements of a standard error-driven learning algorithm such as backpropagation, which requires the use of a continuous differentiable function. To a first approximation, however, this issue can be resolved by using a sigmoidal function for the lower half of the unit input-output function and a

linear function for the upper half. More formally, a unit's activation, a, as a function of its input, i, would be:

$$a(i) = \begin{cases} \frac{1}{1+e^{-i}}, & \text{if } i <= 0 \\ 0.25i, & \text{if } i > 0 \end{cases}$$

This equation approximates a threshold linear function as a continuous transition from a relatively low and stable activation state regardless of the specific amount of input, to a state wherein the activation of the unit varies linearly as a function of the input (the 0.25 value was selected because it is the derivative of the sigmoid for $i = 0$; the equation is therefore continous and differentiable despite being defined in two parts). For ease of reference, the later discussion labels this as a 'siglinear' activation function because its lower half is sigmoidal and its upper half is linear.

The second issue with this adjustment to the formalism is that it would mean not using exactly the same formalism that was used in the biologically-plausible simulation of the settling dynamics account. Hence, a direct amalgamation of these two frameworks would be somewhat heterogeneous. For present purposes, however, this small variation is considered to be a useful exploration of the suitability of this input-output function in the service of increasing the biological plausibility of the framework in a way that appears to be useful for capturing additional effects. Insofar as this work is successful in this first exploration, the earlier simulations could be revisited to assess how well this formalism performs in other contexts.

9.2 Numerosity Judgement Simulation

To assess the degree to which a connectionist model that employs a more biologically-plausible formalism shows the same overall and trial-by-trial effects as human participants, a simulation analogue of the behavioral experiment described in the previous chapter was developed.

Methods

Simulated participants. Two 'simulated participants' completed each of the four main conditions in the numerosity judgment experiment described in the previous chapter. The reduced total number of participants was employed because the simulations were not expected to be influenced by many of the random sources of variability that influence human participants (e.g., fatigue, boredom). Each simulation was initialized with a different set of weights and was presented with a different ordering of the experimental stimuli, the latter being accomplished using the same randomization methods used in the behavioral experiment.

Network architecture. The model architecture is presented in Figure 9.1. The network was composed of 200 visual input units, four excitatory hidden units, two excitatory response units coding 'low' and 'high' responses, respectively, and one inhibitory unit for each of the hidden and response pools. The visual input units were divided into two groups of 100 units. The first group coded for the presence of an asterisk in a particular location in the 10×10 array, whereas the second coded for the complement of this (i.e., for spaces in the array at which an asterisk was not presented). The motivation for including these two sub-groups of visual input units was to approximate how center-on surround-off and center-off surround-on neurons in cortical regions associated with early visual processing (Hubel & Wiesel, 1962); this coding scheme also normalizes the overall amount of activity in the input so that there is not more raw input when more asterisks are presented.[1] Nevertheless, it was intended only to be a very basic approximation of early visual processing because the goal of the simulation was not to simulate the visual processing per se, but only how it could drive the response selection system. The inclusion of hidden units in this simulation was aimed primarily at demonstrating that the

[1] Additional simulations that did not include the second sub-group of visual input units were, nevertheless, successful at learning to perform the task. Inspection of the hidden and output unit dynamics suggested that in these simulations the network learned a hidden representation that was approximately equivalent to the presence of the second sub-group in the main simulation.

biologically-plausible connectionist architecture could successfully simulate the task by mapping through an intermediate pool of hidden units; these units were not strictly necessary for learning the task and similar results were obtained in additional simulations that mapped the input units directly to the output units.

The units in the network were connected as follows: The input units sent excitatory connections to the excitatory hidden units, which in turn sent excitatory connections to the excitatory response units. Additionally, the excitatory hidden units sent connections to all of the other excitatory hidden units (except themselves) and to the inhibitory unit; similar connectivity was also present in the pool of response units. Finally, the inhibitory units sent inhibitory connections to the same units from which they received excitatory connections.

The mean initial weight for all the excitatory connections was set to 0.15 for simplicity after pilot simulations did not show a substantial effect of employing smaller between-layer weights on overall performance. The mean weight value for the inhibitory connections was set to -0.4. Additionally, all of the units received a bias connection with a mean value of -2.19. The exact value of these weights was sampled from a uniform distribution with a range of 1.0 above and below the mean, with the condition that excitatory weights could not have values below 0 and inhibitory weights could not have values above 0. These initial weight values were selected so that all of the units in the network had low activation values at the onset of training. This reflects how before the experiments, participants would not have associated the presentation of different numbers of asterisks with specific behavioral responses.

Additionally, to avoid the network needing to learn positional invariance (i.e., that an asterisk present in one position in the array should, in principle, contribute the same amount of excitation to the hidden layer as an asterisk present in another position in the array), which requires extensive additional training and was not the primary focus of the present simulation, two other minor modifications were made. First, all of the weights from each sub-group of visual input units were constrained to have the same value so that every input unit from each of these sub-

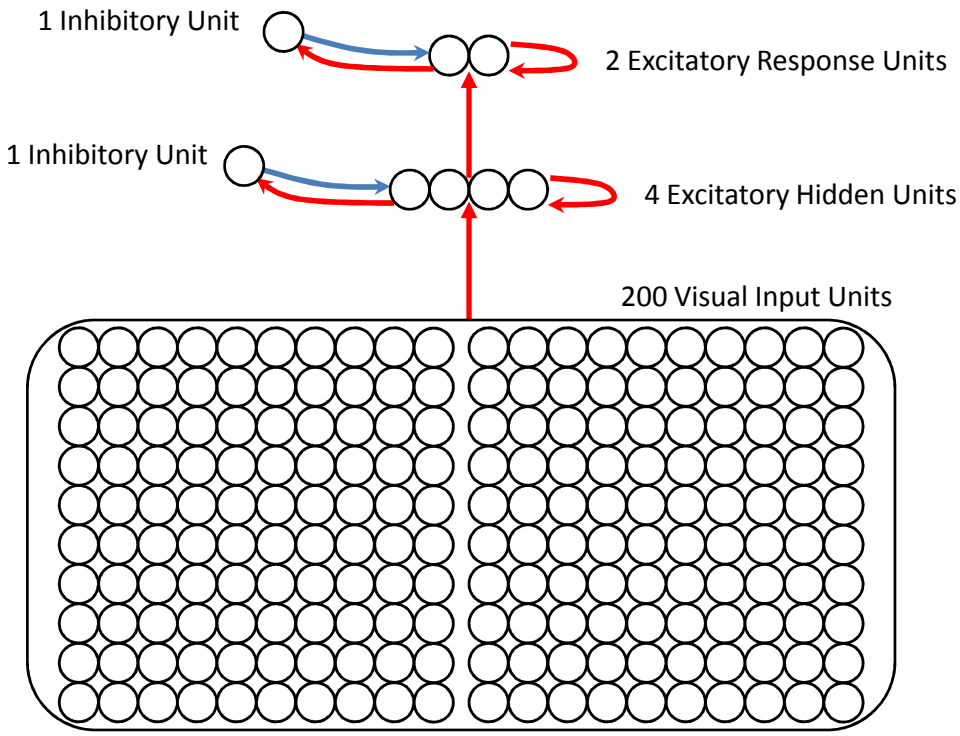


Figure 9.1: Architecture for the network from the numerosity judgement simulation. Red arrows indicate positive (excitatory) connections and blue arrows indicate negative (inhibitory) connections.

groups sent the same amount of activation to the hidden layer. Second, the total output from each sub-group of input units was normalized so that the minimum total output for each sub-group was 0 and the maximum total output was 1 for the most extreme numbers of asterisks that could be presented. This last manipulation was useful for two main reasons: First, it reduced the amount of excitation that was sent to the hidden layer so that the hidden units were not driven too strongly by the visual input prior to training the network to make numerosity judgments. This approximates how, outside of the numerosity judgment tasks, perceiving different numbers of asterisks does not generally cause motor responses. Second, taken together these manipulations were equivalent to employing two input units with activations normalized to the [0,1] range to represent the presentation of different numbers of asterisks in the 10×10 input arrays. This last property allowed for the simplification of certain aspects of the instantiation of the simulation for computational convenience.

All of the hidden and response units integrated their inputs over time (dt = 0.2). A unit's output was a siglinear function of its net input, plus a random amount of normally distributed noise. For all but the input units, the noise had a standard deviation of 0.025 above and below the computed output. For the input units, noise was added so that variability in the total output of the each sub-group had a standard deviation of 0.1. This last manipulation was intended to stand both for processing noise in the input neurons and uncertainty in the identification and representation of particular numerosity values.

Representations. The input patterns for the network were generated by the same sampling procedure that was used for the behavioral experiment. The target output values for the response units were set such that a 'low' number of asterisks had a target of 1.0 for the 'low' unit and 0.0 for the 'high' unit; the complementary target pattern was used for presentations of 'high' numbers of asterisks. These patterns were divided into two groups: pre-training patterns and task-simulation patterns; 1,200 patterns of each type were generated. The number of pre-training patterns was determined in pilot simulations that assessed when, on average, the network reached

an overall rate of accurate responding between 65% and 75%, which was similar to the initial accuracy values that were observed at the beginning of the behavioral experiment. Subsequent to the pre-training, the task-simulation trials were presented.

Pre-training and task simulation. The method of presentation of the pre-training and the task-simulation trials was identical, and the pre-training trials were included only to simulate, in approximate terms, how well the human participants pre-configured their responses system via the task instructions and their existing knowledge of numerosity (c.f. Ratcliff et al., 1999). On each trial, the net input and output of the hidden and response units were set to -2.19 and 0.1 respectively (-2.19 being the net input that corresponds to an output of 0.1 on a siglinear activation function). The network was then presented with the input pattern and trained for 100 unit updates. Error was calculated for the last 95 unit updates. The error was scaled by a factor of 3.0 for the units that were supposed to be off. Taken together, the extended time during which error was accumulated and the error scaling were intended to cause the network to initially learn to activate only the correct output after many updates and then gradually learn to activate the correct output more quickly, just as was observed in the human participants. A unit's target activation was adjusted such that it was considered to be correct once it was either below 0.1 or above 0.9 for units that were supposed to be off and on, respectively. Error was calculated using a two-piece error function such that cross-entropy error was used for activation values below 0.5 and sum-squared error was used for activation values above 0.5; these different error functions have been shown previously to be the appropriate error functions for sigmoidal and linear input-output functions, respectively (Hinton, 1989). Weights were adjusted after each trial using a steepest gradient descent algorithm and a learning rate of 0.01.

Units were considered to have made a response once the activation of one of the response units exceeded 0.5. A fixed as opposed to a dynamic response threshold was employed in contrast to some past work (Armstrong et al., 2009) because this appears to be more consistent with the underlying neurobiology (Hanes & Schall, 1996). If the activation of both both response units

exceeded the response threshold on the same unit update, the unit with the highest activation was considered to have responded. The number of unit updates prior to the response corresponded to the network's response latency. The total length of presentation for each stimulus was, nevertheless, the same for all trials to reflect how stimulus processing (and corresponding error adjustments) may continue even after a response has been initiated (Joordens et al., 2009).

Results

A similar set of analyses to those run on the behavioral data was conducted on the simulation data, although these were simplified in many respects because the simulated participants were not subject to the same biases, either introduced experimentally or otherwise (e.g., post-block feedback, faster responses from the dominant hand), or to additional random sources of variability (e.g., slower responses because of a random lapse in attention). The data screening eliminated all trials for which the model failed to respond. These corresponded to 0.44% of the total data set. No overall statistical screening was applied for the average performance of the simulations because the simulations were necessarily attending to the task instructions and were not subject to as many potential biases as the human participants; there was, however, an implicit upper limit on latencies because trials were run for a maximum of 100 unit updates.

Overall performance in the constant threshold conditions. The overall accuracy and latency (in unit updates) for correct responses for the simulated participants was 92% (SE = 0.2%) and 12.65 updates (SE = 0.08 updates), respectively. In terms of accuracy, this level of overall performance was therefore similar, but slightly higher, than that observed in the human participants (84%), and suggests that the learning rate in the simulation may have been slightly too high.

Figure 9.2 and Figure 9.3 display the accuracy and correct and incorrect latency data, respectively, as a function of the absolute distance in number of asterisks from the response threshold. Both of these plots show that performance improved towards an asymptotic state as the distance

from the response threshold increased, much as was observed in the behavioral data, although performance did increase more sharply as distance from the response threshold increased in the case of the simulations. The latencies for correct and incorrect responses were quite similar up to a distance of approximately 10 asterisks from the response threshold; after that point, the model responded perfectly accurately to all stimuli and so the trace for latencies for incorrect trials is truncated at that point. Consequently, much as for the behavioral data, the informativeness of the error latencies is quite limited, particularly for distances greater than five from the response threshold. Figure 9.4 presents a combined latency-accuracy plot based on the data from Figures 9.2 and 9.3. Interestingly, and in contrast to the separate plots of accuracy and latency, this plot shows that latency differentially increased rather than asymptoted for correct responses as performance on the accuracy measure approached ceiling — just as was observed in the behavioral data. Again, limited weight should be placed on the data associated with incorrect responses because it is based on a relatively small number of trials.

The distribution of correct and incorrect latencies was skewed to the right, as illustrated in the density plots presented in Figure 9.5, and on average latencies were slower and more spread out for the incorrect responses. This parallels the behavioral results. An examination of skew as a function of distance from the threshold, which is presented in Figure 9.6, did not, however, agree with the behavioral results, and showed that skewness decreased as a function of distance from the response threshold. This issue is discussed in more detail in a later section.

In additional analyses not included here for the sake of brevity, it was also confirmed that the latency-accuracy and hazard functions for the present data set were qualitatively similar to those from the analyses of the behavioral data.

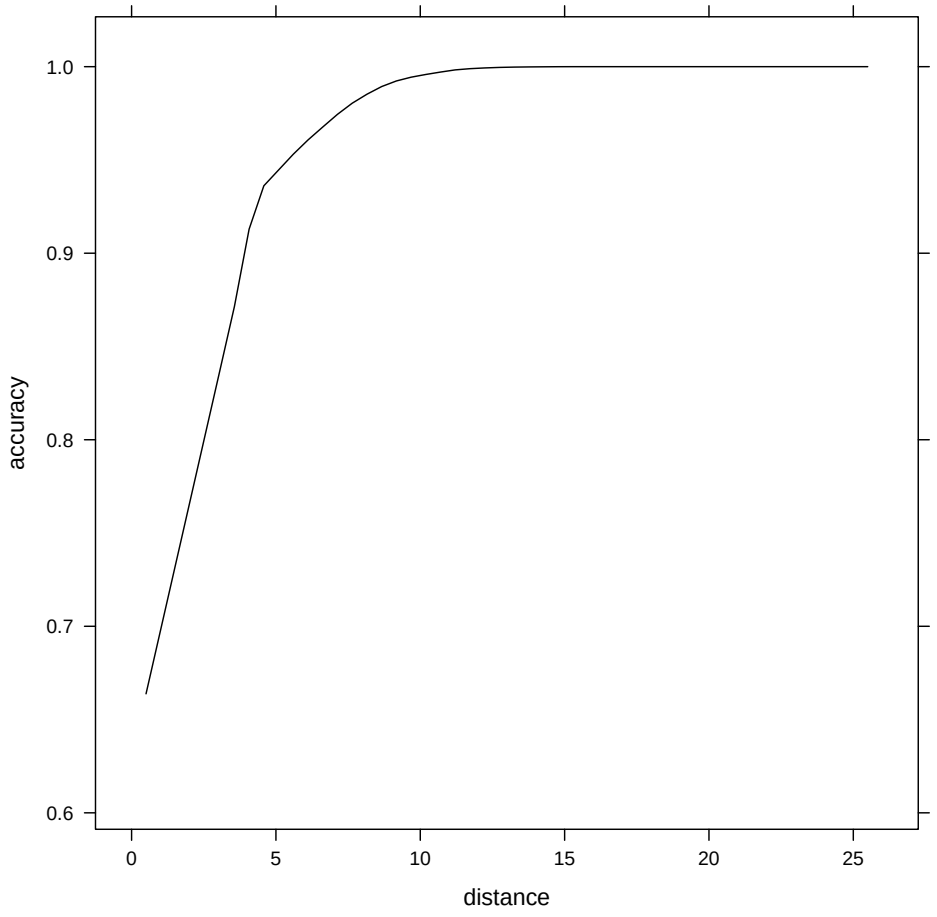


Figure 9.2: Plot of accuracy as a function of the distance in number of asterisks from the response threshold.

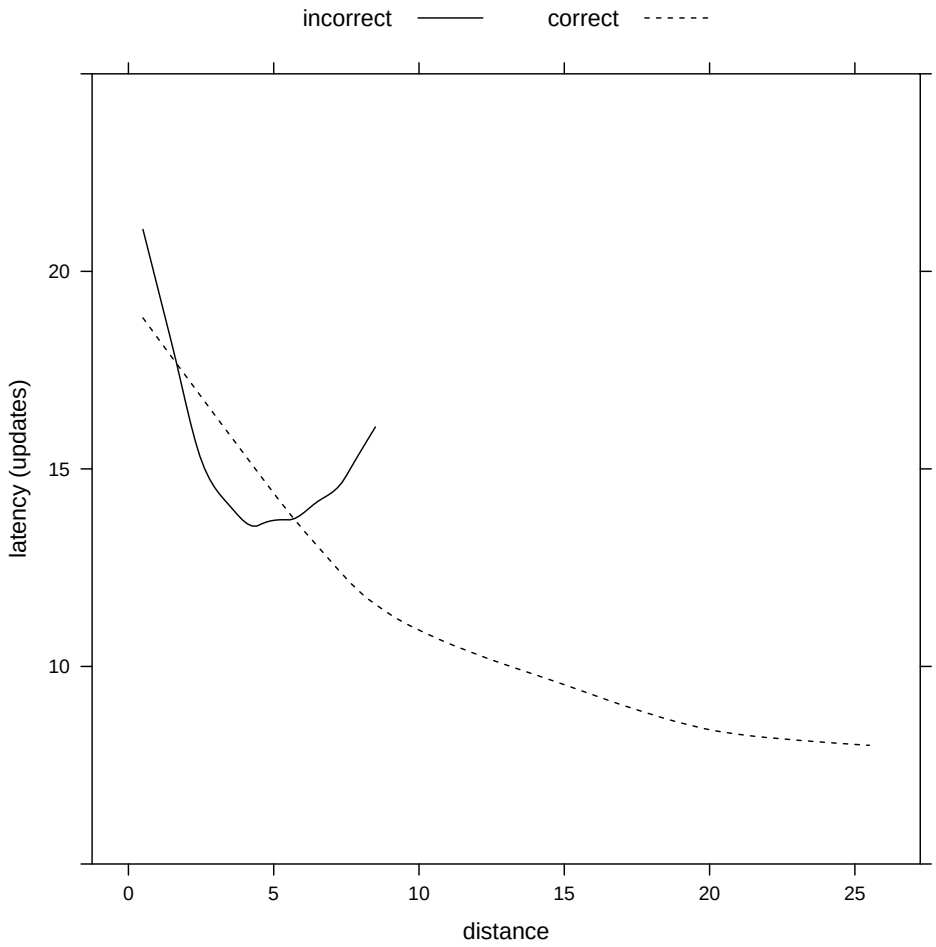


Figure 9.3: Plot of latency as a function of the distance in number of asterisks from the response threshold. Separate lines show the data for correct and incorrect responses.

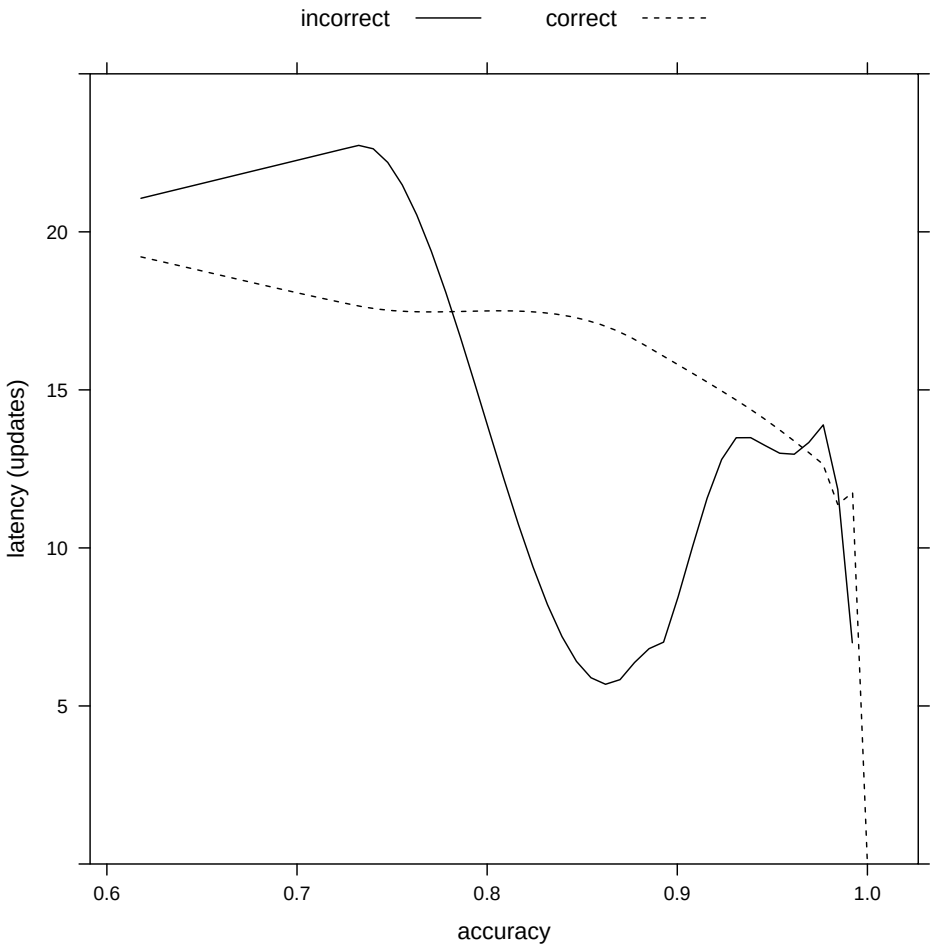


Figure 9.4: Plot of latency as a function of accuracy generated by combining the results from Figures 9.2 and 9.3. Separate lines show the data for correct and incorrect responses.

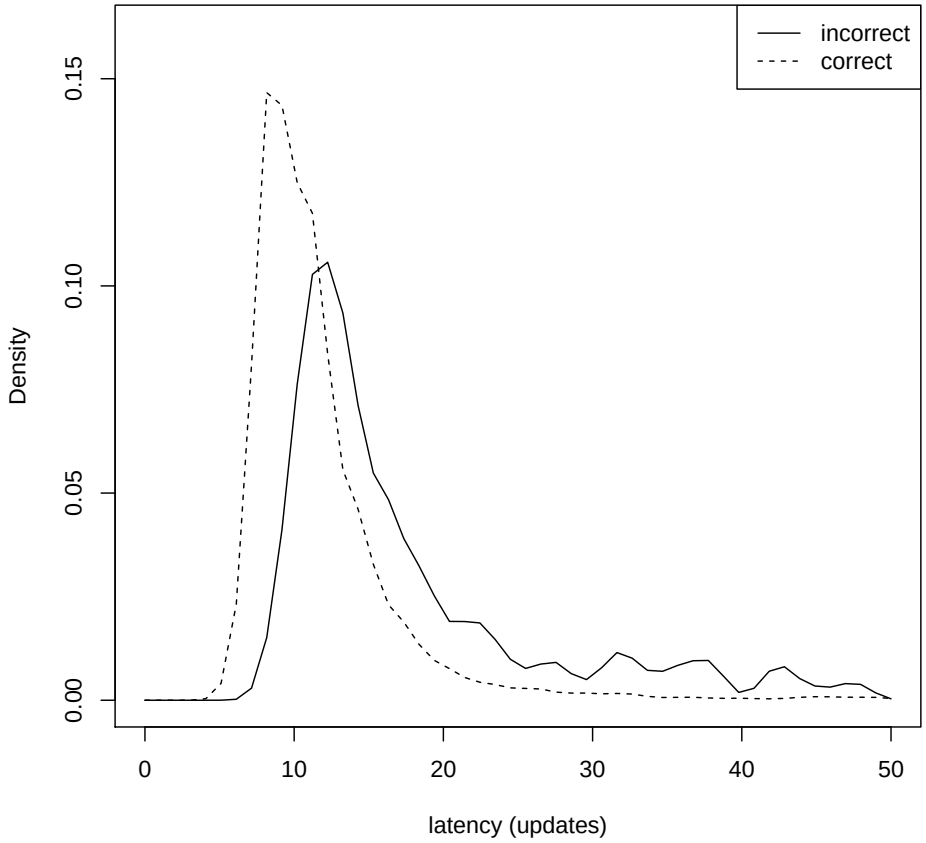


Figure 9.5: Distribution of latencies for correct and incorrect responses.

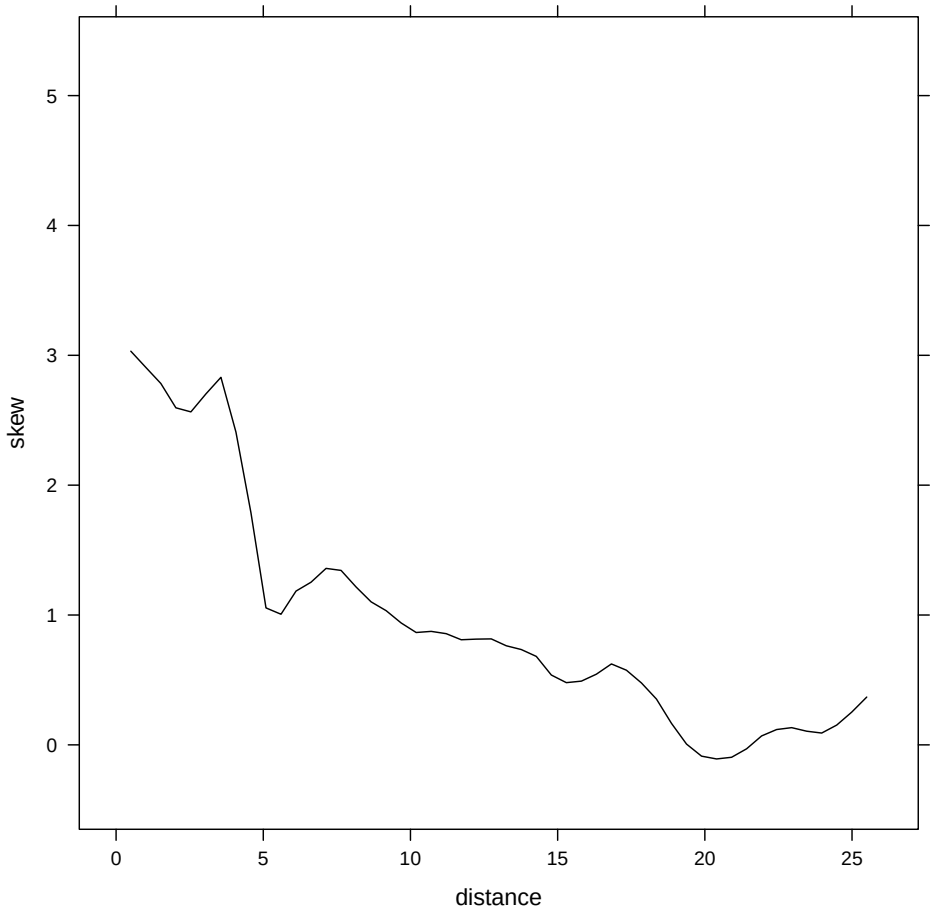


Figure 9.6: Degree of skew in the distribution of latencies for correct responses as a function of distance in number of asterisks from the threshold.

Sequential effects. The trial-by-trial sequential effects in the constant threshold condition were examined in several ways to evaluate how performance changed as a function of previous experience. First, overall accuracy and latency for correct responses were examined as a function of trial rank in Figure 9.7 and Figure 9.8; for the moment, the discussion primarily focuses on the constant threshold condition (same–same). In the accuracy data, the high overall accuracy levels produced instabilities when local regression smoothing was applied to all of the conditions in which a threshold change occured between mega-blocks (a common problem when relatively small sets of binary data are smoothed in this fashion). Visual inspection of the raw accuracy for each trial did, however, indicate that accuracy decreased immediately following a threshold change. The constant threshold condition did show near-asymptotic but slightly improving overall accuracy as a function of additional experience with the task, just as in the behavioral data. In contrast to the accuracy data, the latency data for this condition did, however, show performance improvements as a function of additional experience. This is consistent with the performance observed in the behavioural experiment. These data also did not show an initial increase in accuracy after the first block of trials was presented, which supports the hypothesis that the observation of such an increase in the behavioral data was driven by participants attending to the post-block feedback in the behavioral condition, which was not provided in the simulation. Note also that these sequential effects appear to be present across a range of initial speed-accuracy trade-offs and initial levels of proficiency that resulted from employing different simulated participants in the different conditions.

To gain further insight into how adaptation occurs in the system, additional analyses examined how performance changed as a function of experience for stimuli of varying distance from the threshold. To increase the reliability and simplicity of these plots, stimuli of similar distance from the threshold were binned together each time the distance from the threshold increased by 5 (i.e., items were binned into groups of 1–4 asterisks distant from the response threshold, 5–9 asterisks distant from the response threshold, etc.), and the data were averaged at the block as op-

posed to the trial level. Figures 9.9 and 9.10 plot the results for the accuracy and the latency data for correct trials, respectively. In the case of the accuracy data, only the three groups that were closest to the response threshold were plotted because all subsequent groups showed near-ceiling perfomance and were not differentiable on the plot. For both the accuracy and the latency data, performance generally continues to improve throughout training, although to a lesser extent in the accuracy data, which also reach ceiling performance for many of the groups. The latency data also appear to show the greatest improvements as a function of practice for the groups that are furthest from the threshold. These results parallel the behavioral data and suggest that trial-by-trial adaptations, despite being made to reduce errors, nevertheless primarily lead to decreases in latency in an explicit model of the task.

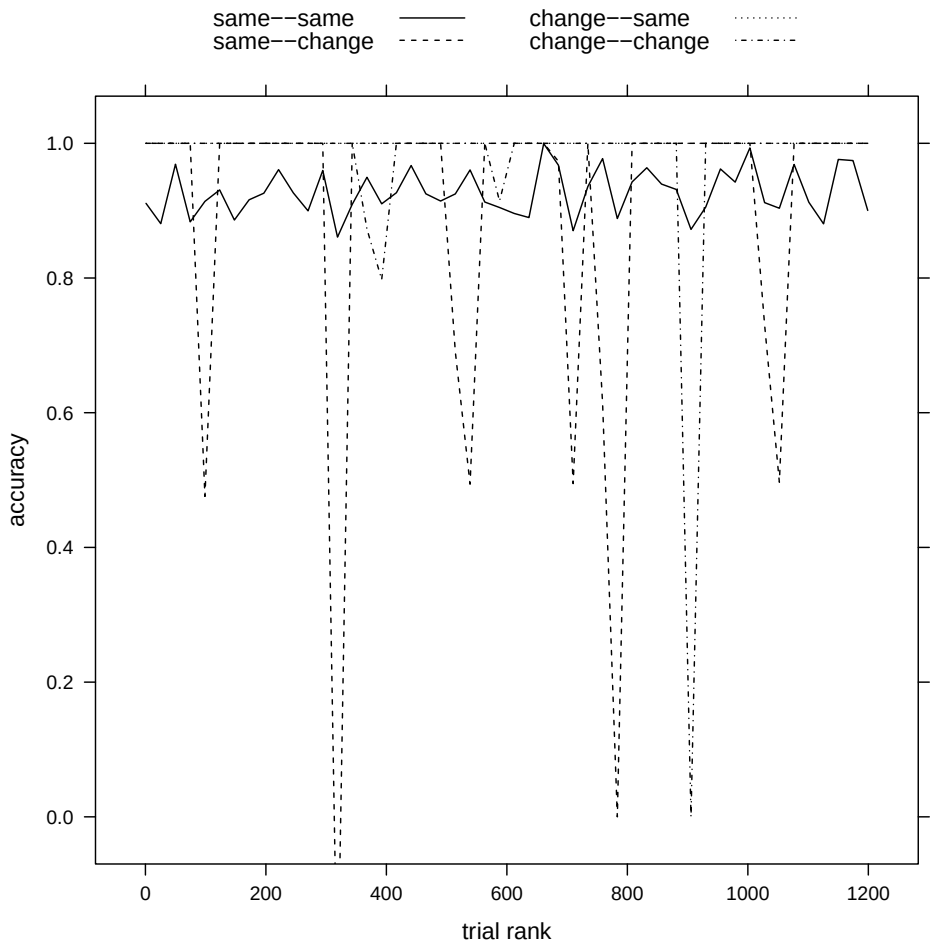


Figure 9.7: Accuracy as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks.

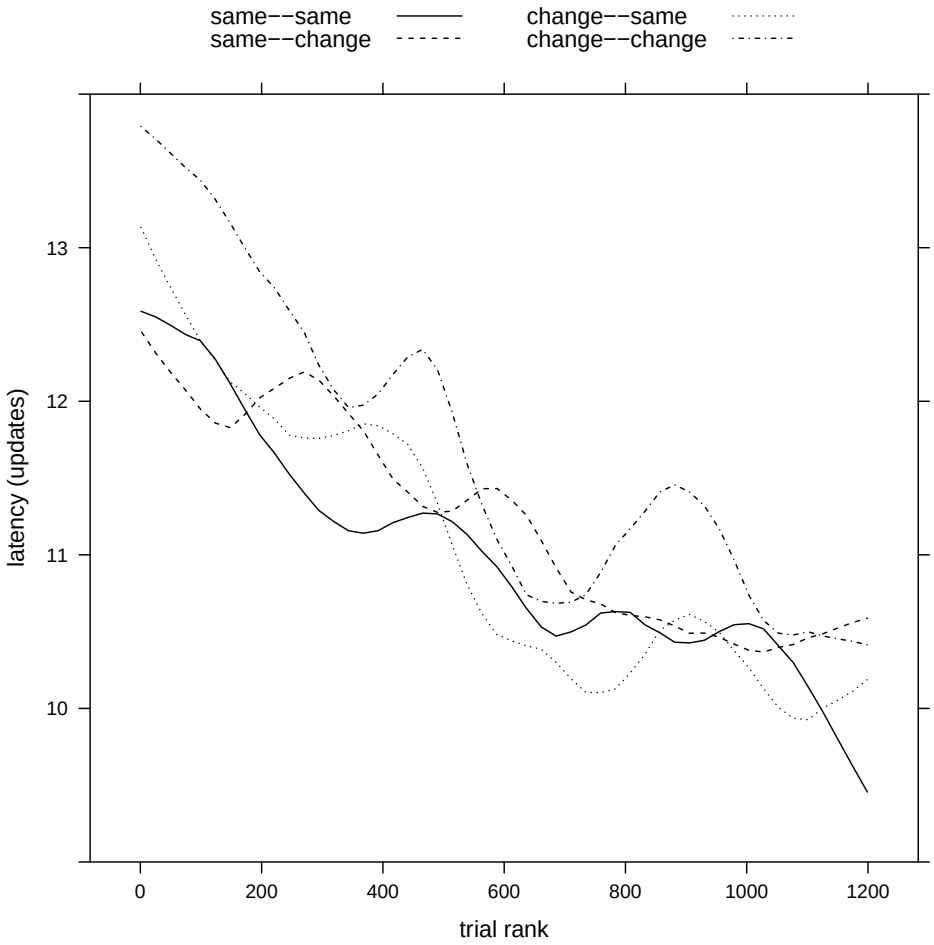


Figure 9.8: Latency for correct responses as a function of trial rank. Separate lines plot the results for different combinations of response thresholds across the mega-blocks.

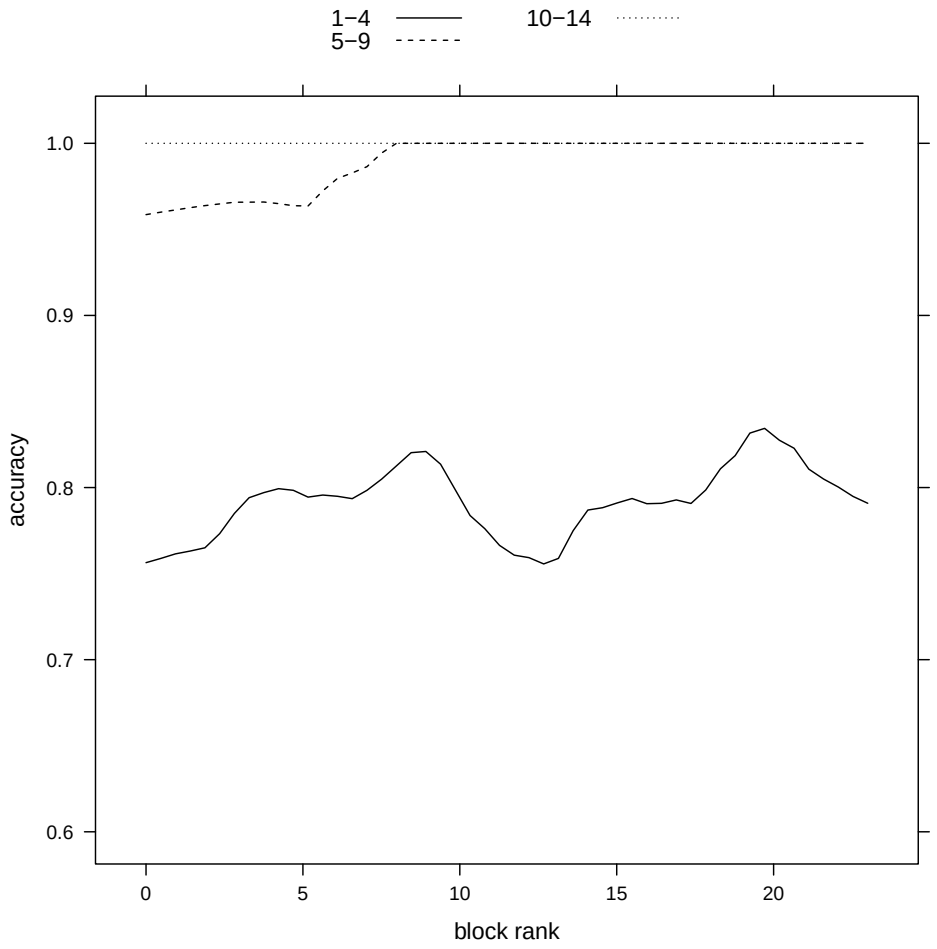


Figure 9.9: Accuracy as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines.

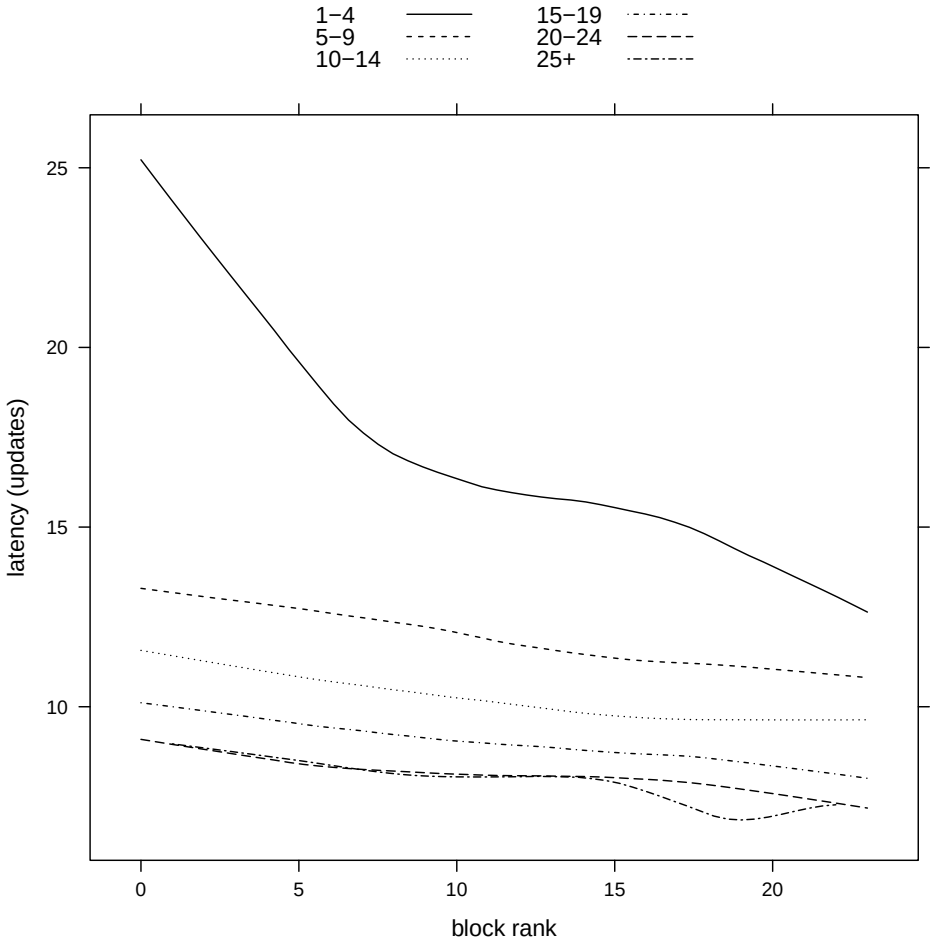


Figure 9.10: Latency for correct responses as a function of block rank. The data for groups of asterisks for a range of different distances from the response threshold are plotted as separate lines.

Having established that the simulations do show adaptive effects as a function of experience with the task, further analyses evaluated which characteristics of the task environment influenced performance and which characteristics were predictive of trial-by-trial sequential effects. A mixed-effect regression model (Baayen et al., 2008) was used to quantify the effects of several factors on response accuracy and on the latency of correct responses. The details of these analyses were identical to those from the behavioral analyses, with the exception that a number of variables that were included in the participant analyses were dropped because they were irrelevant (e.g., the button used to signal high responses), not part of the simulation (e.g., post-block feedback), or were considered unnecessarily complicated for assessing initial model performance (e.g., only variables related to the immediately preceding trial were included instead of variables related to the two preceding trials).

Accuracy. The coefficients for the predictors of the accuracy data are presented in Table 9.1. Overall, the similarity between the significant predictors in the behavioral analyses and in the simulation analyses were quite high with respect to the properties of the target trial: In both cases the more frequently presented stimulus was responded to more accurately, stimuli that were further from the response threshold were responded to more accurately, and the amount of experience, in number of trials, did not strongly predict subsequent accuracy. However, the analyses were not in very strong agreement with respect to the predictors from the previous trials, although in both cases the analyses showed weak trends for overall accuracies as a function of additional experience. Some of these disagreements are weak, for instance, concerning how well the distance or accuracy of the response related to the preceding stimulus predicted the accuracy of the subsequent trial. For both the behavioral and simulation analyses, the statistical test identified significant predictors that were nonsignificant in the behavioral analyses, although whether this was simply the result of increased variability in the behavioral data is unclear. More troubling, however, were the predictions related to the latency of the previous trial and whether the current trial consisted of a repetition of the same stimulus or response that was made on the

previous trial. Whereas the behavioral data indicated that slower preceding responses should increase accuracy, repetitions of the stimulus should increase accuracy, and repetitions of the response should decrease accuracy, the model did not show a significant effect of the latency of the preceding trial and showed significant effects in the opposite direction for the other two predictors. Additional discussion of these discrepancies is deferred to a later section.

Latency for correct responses. The coefficients for the predictors of the latency data for correct responses are presented in Table 9.2. Again, the predictors related to the properties of the target trial being analyzed are identical for both the simulation and the behavioral data: more frequent responses are made more rapidly, increased distance from the response threshold leads to more rapid responses, and more experience with the task leads to faster responses.

There are, however, disagreements between the significance patterns associated with predictors that reflect performance on previous trials, although in both analyses additional experience with the task predicted faster latencies. For instance, the distance of the previous trial from the response threshold was not a significant predictor in the simulation, but was a significant predictor in the behavioral data. The accuracy of the previous trial was a nonsignificant predictor in the simulation data but higher accuracies on preceding trials predicted decreased latencies on subsequent trials in the behavioral data. The behavioral and simulation data did, however, agree with respect to the influence of the previous trial's latency on the latency of a subsequent trial, and agreed in sign if not in absolute levels of significance that repetitions of a stimulus lead to faster responses.

Table 9.1: Statistics for the predictors in the accuracy analysis

	β	SE	*z*	*p*	
(Intercept)	0.9959	0.6898	1.4	0.149	+
fstim	−1.0768	0.2366	−4.6	0.000	***
distance	0.7174	0.0654	11.0	0.000	****
trial	0.0003	0.0003	1.0	0.319	
distance_1	0.0871	0.0215	4.1	0.000	***
acc_1	−1.9869	0.4651	−4.3	0.000	***
rt_1	−0.0084	0.0131	−0.6	0.524	
repfstim_1	−1.5705	0.3666	−4.3	0.000	***
represp_1	5.6819	0.6916	8.2	0.000	***

Note. + = p < .15, * = p < .05, ** = p < .01, *** = p < .001.

Table 9.2: Statistics for the predictors in the analysis of the latencies for correct trials

	β	SE	*t*	*p*(mcmc)	
(Intercept)	16.99	0.96	17.8	0.030	*
fstim	2.55	0.27	9.3	0.000	***
distance	−0.43	0.02	−22.4	0.000	***
trial†	0.00	0.00	−9.8	0.000	***
distance_1	0.01	0.02	0.4	0.699	
acc_1	0.72	0.70	1.0	0.302	
rt_1	0.03	0.02	1.5	0.141	+
repfstim_1	−2.24	0.91	−2.5	0.012	*
represp_1	1.30	0.96	1.4	0.182	

Note. + = p < .15, * = p < .05, ** = p < .01, *** = p < .001.

† The sign of the trial predictor was negative but rounds to zero.

Adaptive effects. The overall effect of practice for the constant threshold condition and each of the variable threshold conditions is plotted in Figure 9.7 and Figure 9.8 for accuracy and latency for correct trials, respectively. These figures show that despite the considerable variability within the conditions and smoothing in the accuracy data, there is at least some basic evidence that accuracy decreases and latency increases following a threshold change, although this was more pronounced in the conditions in which the threshold changed multiple times (change–change) or changed early (change–same), with little evidence for substantial overall effects of threshold changes when they were presented later (same–change).

To evaluate whether different portions of the distribution were more affected by the change in threshold than others, the performance in each mega-block for each condition was plotted relative to a standardized response threshold (see the analogous analyses in the previous chapter for details). These plots are presented in Figure 9.11 and Figure 9.12 for accuracy and correct latency, respectively. Once again, the local regression smoothing suffers some difficulty with the accuracy data, although these are primarily limited to the first mega-block when performance is more variable and below its later near-asymptotic level. These plots show that although only the correct responses for the central portion of the asterisk distributions were affected by the change in threshold, performance across a range of values was altered following a threshold change. Any threshold change also resulted in at least a small overall impairment relative to the constant threshold data, particularly in terms of latency. Furthermore, these data show that adaptation occurs at approximately the same rate regardless of whether the threshold changes early or late, although adaptation may be slightly slower in the latency data following two threshold changes. Thus, these data also parallel the results observed in the behavioral experiment in general terms.

To be consistent with the behavioral analyses, Figure 9.13 plots correct latency as a function of trial rank for different standardized numbers of asterisks and combinations of constant or changing response thresholds across mega-blocks to provide a more detailed presentation of the adaptive effects. The accuracy data plots are omitted because of singularity issues with the

local regression, although similar effects appear present in a visual inspection of the raw data. The data from three groups of asterisks are presented: one group fell just below the initial low threshold (37–45), one group fell just above the initial low threshold and just below a later high threshold in conditions in which a threshold change occured later (46–54), and one group was well above the initial low threshold but was immediately above a high threshold if the threshold value changed (55–63). Data from the most extreme standardized numbers of asterisks were excluded because the rarity of these trials for any given trial rank was such that reliable curves could not be achieved without considerable smoothing. These results show that adaptation occurs over a relatively extended period of time following a change in the threshold, just as was observed in the behavioral data, although adaptation may be somewhat faster in the simulation. This may simply be due to using a learning rate that was slightly too high. The magnitude of these changes was also larger for the two groups that always received consistent feedback and for which responses were typically quite accurate relative to the group for which the correct response changed across mega-blocks and accuracy was typically quite low—once again, indicating a reasonable similarity between the simulation and behavioral data.

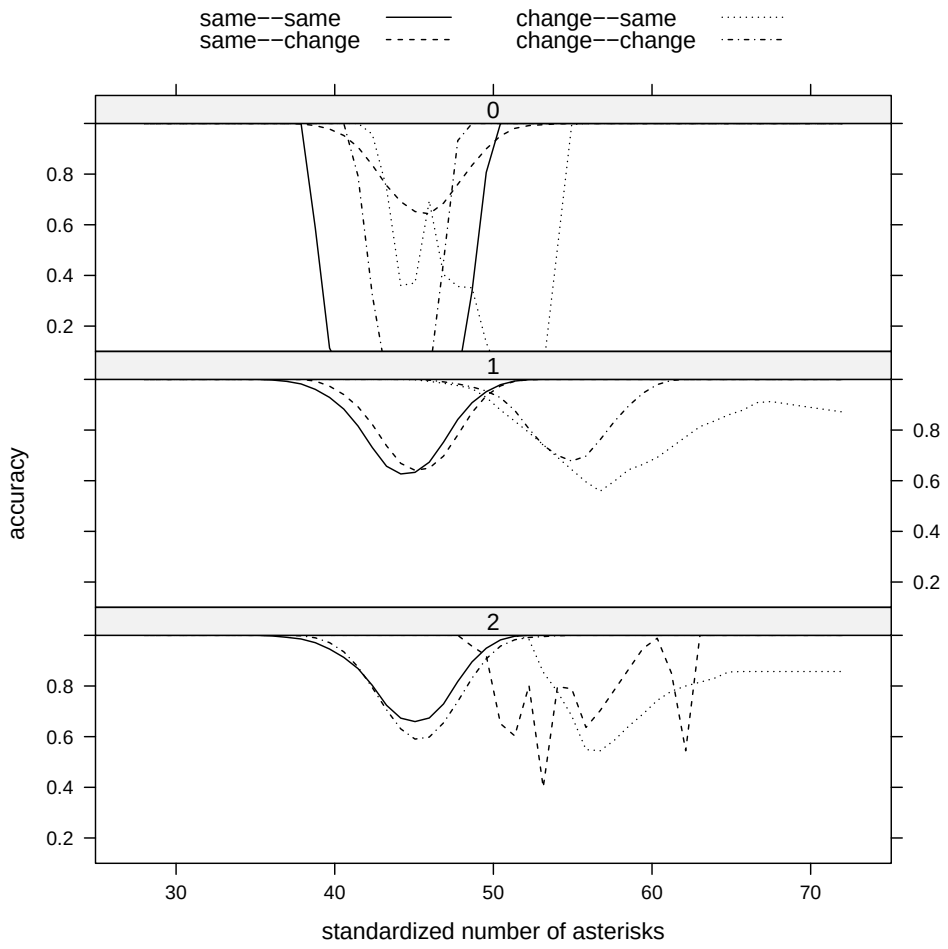


Figure 9.11: Accuracy as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions.

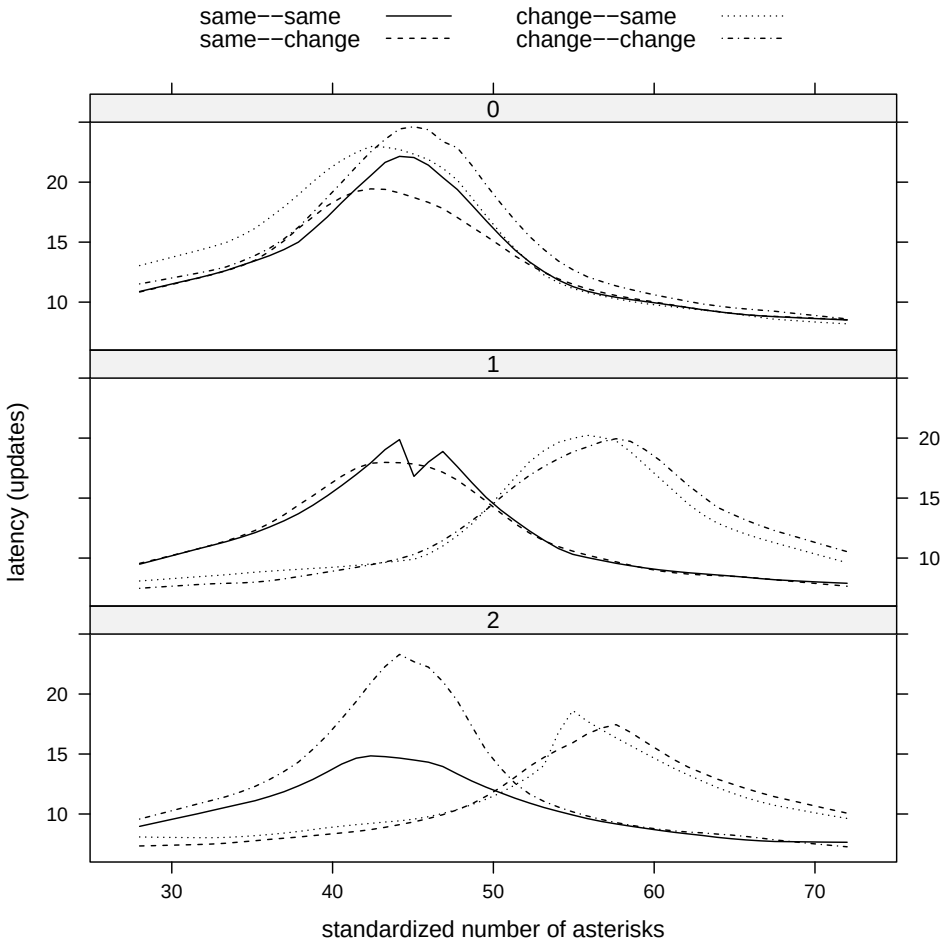


Figure 9.12: Latency for correct responses as a function of the standardized number of asterisks. Each panel represents average performance in one of the three mega-blocks and separate lines plot the results from each of the different conditions.

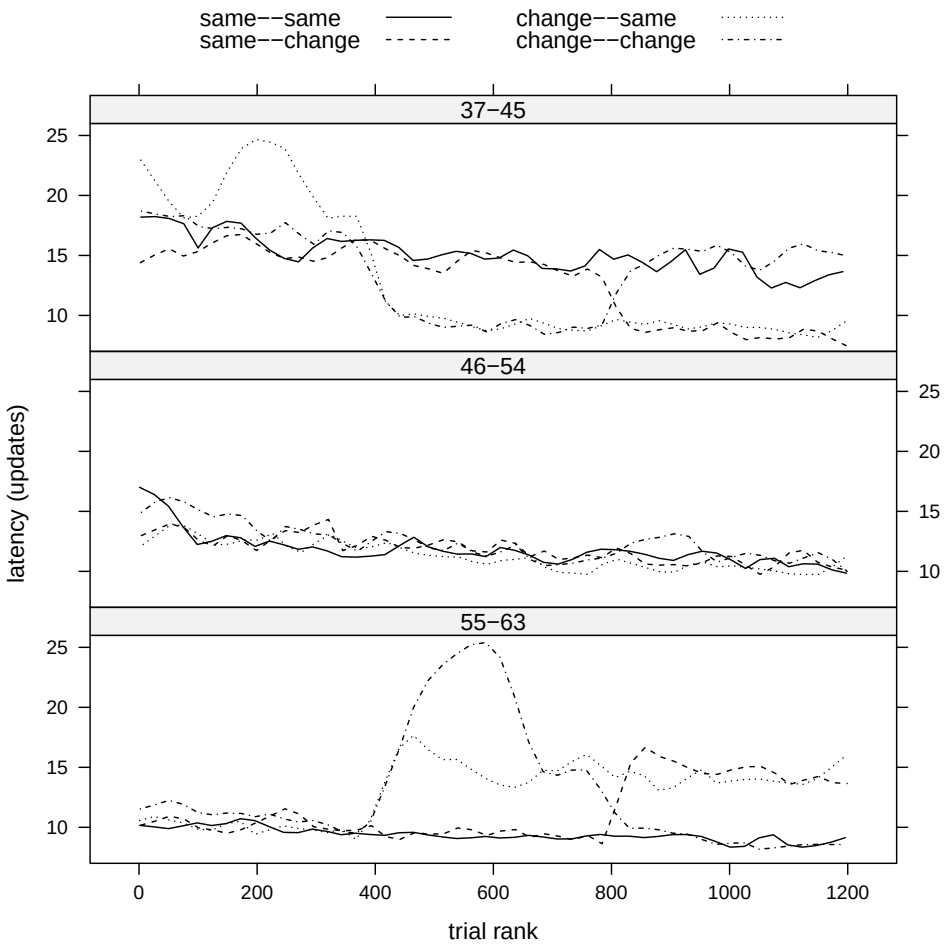


Figure 9.13: Latency for correct trials as a function of trial rank. Separate panels present different groupings of standardized number of asterisks and separate lines within a plot represent the different conditions.

Activation trajectories. In the course of running the simulations, an additional emergent property of this architecture was observed: Despite employing a unit input-output function for which there is no explicit upper bound, the units tended to settle to stable asymptotic activation levels. This was true both if the input corresponded to a number of asterisks that was near or far from the response threshold, although the asymptotic level of activation did differ across these two cases. This property is of interest because the leaky accumulator model needed to add an explicit leak current to avoid runaway activation if a stimulus was presented for an extended period of time. The present work, however, suggests that this is an emergent characteristic of a more biologically-plausible connectionist formalism (although obviously, more work will be needed to assess the generality of this observation). Examples of the activation trajectories for the two response units for the presentation of two stimuli that correctly activate the same response unit but differ in terms of their distance from the response threshold are presented in Figure 9.14.

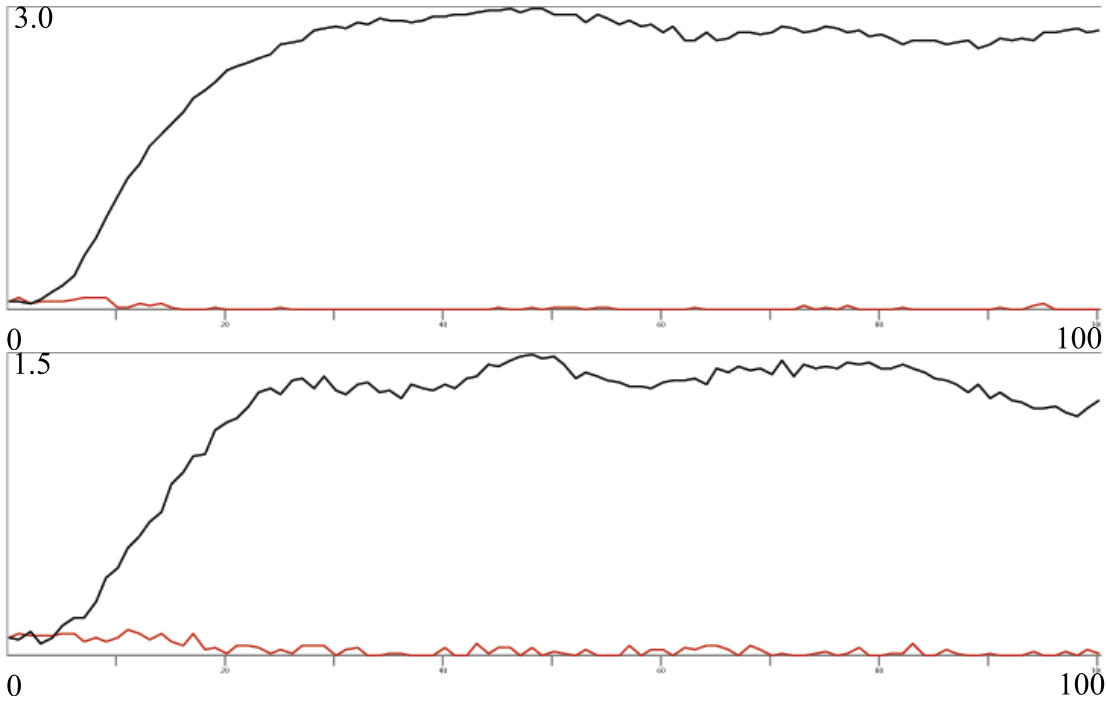


Figure 9.14: Activation trajectories for the 'low' (black) and 'high' (red) response units. Both plots show the correct activation of the 'low' response unit following the presentation of a 'low' number of asterisks. The top panel shows the trajectories associated with the presentation of a stimulus that was far from the response threshold (32 asterisks). The bottom panel shows the trajectories associated with the presentation of a stimulus that was close to the response threshold (44 asterisks). Note that different y-axis scales were used for each plot to highlight the relative similarity between both plots.

Discussion

In contrast to the claims of Ratcliff et al. (2009), the more biologically-plausible connectionist network reported here succeeds quite well at capturing, to a qualitative degree at least, a large portion of the effects that were observed in the behavioral experiment reported in the previous chapter. In terms of overall performance, these included the effects of distance from the threshold on accuracy and latency, latency as a function of accuracy, and the distribution of correct and incorrect latencies. Moreover, this model also succeeded quite well at capturing the trial-by-trial sequential and adaptive effects that are not captured by models that are simply fit to an aggregate data set, as is typically done with the diffusion model. In terms of sequential effects, these included: small increases in accuracy and substantial increases in latency as a function of experience with the task, and small performance improvements in terms of accuracy and more substantial performance improvements in terms of correct latencies for stimuli that were closer to the threshold. Finally, in terms of adaptive effects, these included: weak evidence for decreases in accuracy following a threshold shift and stronger evidence for increases in latency following a threshold shift, approximately equal rates of adaptation following a second threshold shift to the rates of adaptation observed following the first threshold shift, slower adaptation if the threshold shift occured during a later block than an earlier block, and an extended period of adaptation that spanned approximately 100 trials. Additionally, an inspection of the activation trajectories for the decision units suggests that the biologically-plausible connectionist framework employed for this simulation may capture certain asymptotic effects that have previously been addressed in the leaky accumulator model by the addition of an additional parameter (activation leakage). Obviously, more work will be needed to assess the generality of this claim, but insofar as this holds, it will lend further support to the validity and parsimony of the present approach. Moreover, all of this was achieved in a simple model that was not optimized using an unconstrained algorithm to fit the empirical data as is typically done for other models; rather, the initial parameters were estimated on the basis of general connectionist principles and basic assumptions about the initial

configuration of the response system and the visual system (e.g., at the beginning of training, visual input did not drive the response selection system sufficiently to initate a response), as well as the overall performance observed in the behavioral data (e.g., initial and final accurates were approximately 75% and 90%, respectively).

Taken together, these results indicate that at least to a first approximation, the learning, processing, and representation principles of the connectionist framework are a reasonable basis for understanding response selection and other cognitive processes more generally. Moreover, the fact that the model presented here can capture, to a first approximation, the sequential and adaptive effects on the basis of simple error driven-learning principles suggests that this framework may be more suitable for studying response selection both in two-alternative forced-choice tasks and beyond.

Without retracting these general claims, however, the present approach does still leave something to be desired in several respects. The most problematic issues were the effects of several properties of the previous trial on performance on subsequent trials. Here, the results of the simulation were somewhat inconsistent. In broad terms, the simulations did produce highly similar performance to the behavioral data on a large number of metrics, including several overall assessments of sequential and adaptive effects. However, the statistical analyses of the effects of several of the predictors related to the accuracy of the previous trial, as well as to the repetition of the previous stimulus or response, were not well captured by the model. Taken together, these results indicate that although error-driven learning is a reasonable model of many aspects of trial-by-trial learning, additional factors must be considered to capture the finer-grained details of the data. One important factor that the behavioral results indicate may be lacking in the present simulation was a better representation of how the previous trial influences the processing of a subsequent trial. In particular, the only influence that the previous trial could have on a subsequent trial was through the error-driven weight changes that occured at the end of each trial—the outputs of the units were reset between trials. However, the behavioral data indicate

that the representations of the stimulus and of the response may be primed in a fashion that is not captured by these adjustments. A revised version of the simulation could assess this possibility by allowing the effects of the previous trial to influence the subsequent trial not only in terms of weight adjustments but in terms of residual activation. Ultimately, this might be examined best by modeling the inter-trial interval explicitly (as was done by Purcell et al., 2010). However, a simpler first step would be to adjust the initial activations of the units in the network to also average in a small portion of the final activation at the end of the previous trial when the units are reset. This would also bring the present work more in line with other models of activation-based priming (e.g., Plaut & Booth, 2000).

In addition to this more critical failure, the network also failed to capture certain detailed qualitative effects such as how the skewness of the latency distribution varied as a function of the distance from the response threshold. Although in both data sets the distribution was positively skewed, skewness increased as a function of distance from the response threshold in the behavioral data, whereas the opposite was true in the simulation data. Whether this reflects an inaccuracy in the formalism more generally, or represents an emergent characteristic of this particular simulation that simply did not reflect the detailed characteristics of the behavioral performance, is unclear. One possibility that is suggested by other modeling work is that capturing this particular effect may require the introduction of alternative assumptions regarding the distribution of noise in the simulation, as was done in the leaky accumulator model (Usher & McClelland, 2001) and in the diffusion model (Ratcliff, 1978). However, before exploring such a modification it is worth assessing whether this effect emerges in a second set of simulations that are informed by outcome of the co-ordinated computational and behavioral work reported in this dissertation. In particular, these simulations would incorporate the effects of residual activation on subsequent trials, employ a smaller learning rate to better approximate the asymptotic accuracy levels and the rate at which participants adapted in the behavioral experiment, and test a larger number of participants to make the comparisons between the simulations and the human participants more

transparent and less prone to failing the assumptions of the local regression smoothing. Given the principled motivations for these modifications, it is worth exploring whether an emergent characteristic of these changes is an elimination of this problematic result, as well.

Chapter 10

Summary and Discussion of Part II

Ratcliff and his collaborators (Ratcliff & McKoon, 2008; Ratcliff et al., 1999) have argued that the diffusion model is a near-optimal model of response selection. They have also argued that connectionist models are ill-suited for modeling response selection because they do not capture certain key behavioral effects (which at least with respect to trial-by-trial adaptation, are also not captured by the diffusion model, although this is not emphasized in their work). For example, error-driven learning algorithms predict specific changes in performance following incorrect responses that were not observed in the behavioral data from a numerosity judgment task reported by Ratcliff et al. (1999; e.g., repetitions of the same incorrect response should be less likely). Similarly, connectionist models were not able to adapt to changes in the distribution of stimuli as rapidly as human participants, and did not show the same distributions of error rates that participants exhibited. Consequently, important modifications to the connectionist framework would be needed to bring it in line with these data and allow it to capture the same effects that the diffusion model can capture. Ratcliff et al. (1999) anticipated that addressing these failings of the connectionist framework would be a very challenging undertaking and ended their paper by stating that "[they] hope that others are encouraged to explore these possibilities, using [their] data and conclusions as a starting point, with the aim of constructing successful connectionist models" (Ratcliff et al., 1999, p. 296).

Following the publication of Ratcliff et al.'s (1999) paper, considerable work has been devoted to developing a connectionist model of response selection. One major branch of this work has focused on the development of connectionist models that are specialized for simulating response processes with the same degree of precision as the diffusion model. However, to succeed at fitting empirical data with a high degree of precision, these models sacrifice key connectionist principles such as the generality and parsimony of the modeling framework and its integration with a principled theory of learning (e.g., Usher & McClelland, 2001). Given the critical importance of these principles with regards to the domain-general nature of connectionist models and theories, however, this was not viewed as an acceptable trade-off.

Diverging from past work, the numerosity judgement experiment reported here examined whether the original conclusions drawn by Ratcliff et al. (1999) were based on an esoteric and non-representative set of data, which would undermine their claims regarding the general failures of connectionist models. This was done in the context of an extension of their original study which was designed to minimize the likelihood that participants would employ explicit strategies to alter the configuration of their decision system. This study also focused on the effects observed during the first 1,200 trials of experience participants had with the task, rather than the asymptotic performance observed when participants completed 12,000 trials and the data from the first 1,200 trials were discarded.

Contrary to the original claims by Ratcliff et al. (1999) that their results should be broadly representative of a wide range of tasks, these minor changes produced several results that were not consistent with what they had reported, but that are consistent with other studies (e.g., Armstrong et al., 2009), despite producing qualitatively similar results on some metrics. For instance, the extension of their study showed that participants did adjust their performance on the basis of the accuracy of the previous trial to increase their accuracy on the subsequent trial, and that the effects of these adjustments were stronger in the latency data for correct responses (although the effects were significant in the accuracy data as well). Furthermore, the results showed that

participants tended to make smaller adjustments following incorrect responses as a function of increased experience with the task (as expected if there are fewer and less egregious errors to correct). This suggests that error-driven learning principles are, in fact, consistent with how participants adjust their performance when the utility and availability of explicit adaptation strategies is minimized. Moreover, the results showed that participants' adaptation to changes in the distribution of the stimuli were much slower than those reported in the original study. This is consistent with the notion that participants in the original experiment were employing explicit strategies to override or supplement error-driven learning processes so as to make rapid adjustments to their decision system in some instances (e.g., following changes in how frequently different distributions of stimuli were sampled), and minimal adjustments to their decision system in others (e.g., following feedback that a response was incorrect despite it being the optimal response from a statistical standpoint). Finally, the results of the behavioral work indicated that the latency data for incorrect responses were not very reliable in this task and should therefore not be used to make strong claims about whether a connectionist model can successfully simulate response selection, contrary to the claims of Ratcliff and colleagues.

Following up on the behavioral experiment, a connectionist model was implemented using a more biologically-plausible and domain-general connectionist framework to determine whether such a model could serve as a reasonable account of response selection in that task. The results of the simulation showed that, by and large, this was the case: The simulation produced overall performance that was quite similar to that produced by the human participants. Moreover, the simulation also produced trial-by-trial sequential and adaptive effects that were quite similar to those observed in the behavioral work; effects that are typically not addressed by the diffusion model or connectionist models of response processes because those models are fit to aggregate data sets. There were, admittedly, several minor inconsistencies between the simulation data and the behavioral data. However, after a careful consideration of the representations and processes that may account for these effects that were not built into this simulation, there is reason to be

optimistic that these issues can be addressed in future simulation work.

Taken together, this work highlights the many advantages of building models using general principles that can make contact with the representations, processes, and neural mechanisms that determine performance in a given task. It also highlights many of the critical limitations of developing overly-specified models that apply only to particular phenomena within a specific domain and that do not make contact with the neural and computational mechanisms that underlie the behavioral phenomena. For example, whereas the diffusion model is fundamentally a model of two-alternative forced-choice responding and makes no claims beyond that domain (the leaky integrator model being only slightly more general), the present model is only one instance of a broad class of models that can be applied to understand a wide range of domains. Consequently, the development of the model presented in Part II could be informed by much of the work done in investigating more biologically-plausible connectionist networks that was completed in the pursuit of an improved model of semantic settling dynamics, and by connectionist learning principles more generally. Similarly, the successful use of a siglinear activation function in the response selection model can now be extended back to the original settling dynamics work to evaluate the generality of this aspect of the model and its utility in other domains. The ability to make contact with these additional neural constraints was also only possible because the model operates at the level of (simplified and abstracted) neural information processing; the diffusion model is too abstract to make such direct ties and is therefore limited in the degree to which it can make contact with these data. Furthermore, the trial-by-trial adaptive effects that are currently outside the scope of the diffusion model and of the leaky integrator model were well-captured in the present model, which was able to draw on an extensive history and supporting literature concerning error-driven learning that cannot be easily incorporated into those models. These learning principles also provide a basis for understanding how and why particular model parameters emerge as a function of experience with the task that are not provided via an unconstrained parameter fitting algorithm— such an algorithm may provide better quantitative

data fits of particular phenomena, but cannot provide a principled justification for those choices. Consequently, the present model not only appears to be a reasonable model of response selection, in many respects it appears to be a superior model of response selection to the dominant accounts in the field.

The outcome of Part II therefore bodes well for the use of this model of response selection in developing an integrated account of the different semantic ambiguity effects observed in different tasks. Additionally, the added advantages of this framework also promise to facilitate a broader and more flexible theory of response selection. In particular, this framework is suitable for extension to tasks that involve distributed as opposed to localist representations of the response, and to tasks that involve the generation of novel responses (e.g., word and nonword naming, Plaut et al., 1996; Seidenberg & McClelland, 1989). This is critical in light of the risks that are inherent to making strong claims regarding the general suitability of a particular framework as a model of response selection solely on the results associated with a single task, as was done by Ratcliff et al. (1999). Given how relatively minor modifications of an existing task lead to such substantial changes in the observed effects, a better approach would be to explore a wider range of tasks prior to drawing strong conclusions about particular models. To this end, it may be useful to consider a broader set of response selection tasks than those that involve two-alternative forced-choices on very simple perceptual evidence (e.g., lexical decision, semantic categorization, word and nonword naming). In principle, these tasks would appear to share many of the same demands as the perceptual tasks, but each provides unique insights into how the underlying representations drive responses, and of how the responses themselves are represented. Given the success of the connectionist framework at capturing a vast range of different phenomena it seems likely that this framework will continue to prove useful in understanding the data from these tasks. These data are also expected to further undermine the utility and relevance of domain-specific processing models and of the literature's focus on fitting the details of specific tasks without understanding the broader learning, representation, and processing assumptions that underlie the phenomena of

interest.

Part III

Summary and Conclusion

Chapter 11

Summary and Conclusion

Developing a theoretical account of a range of semantic ambiguity effects is challenged by the complex and often contradictory patterns of results in the behavioral data, as well as by claims that such effects cannot emerge in the semantic representations of a parallel distributed processing system. After reviewing the literature, however, the time course of semantic settling dynamics, as conceptualized in a parallel distributed processing system, showed considerable promise for serving as a parsimonous and general means of accounting for a large range of effects. However, this 'settling dynamics account' relied on a complex set of representation and processing assumptions that may not necessarily unfold as described in a system that instantiated the core principles of the theory.

To substantiate the account and demonstrate that semantic settling dynamics could, in fact be the basis for explaining a range of semantic ambiguity effects, several connectionist models of ambiguous word comprehension were developed. These simulations—and in particular, the simulation that employed more biologically-plausible constraints—reproduced the basic patterns of effects that were originally sketched verbally and predicted a novel pattern of effects for a task that required an intermediate amount of semantic processing relative to (easy) lexical decision and (hard) semantic categorization. Future work that improves on the representation of context in these simulations was also discussed.

To further support and evaluate the settling dynamics account and contrast it with the predictions of the decision system account, behavioral studies of lexical decision were conducted to evaluate whether varying task difficulty—and the presumed amount of settling in semantics—while holding task constant would lead to different patterns of effects. The first experiment employed the same word stimuli that were used in a previous study by Rodd et al. (2002) and manipulated task difficulty via the wordlikeness of the nonwords. On first examination, these results supported the predictions of the settling dynamics account, but follow-up analyses and experiments revealed major issues with the stimuli. These issues were addressed by developing new methods for norming relative meaning frequency and for selecting stimuli that were optimized to minimize potential confounds and maximize the magnitude of the effects of interest. A new study that employed these methods and varied difficulty via nonword wordlikeness and stimulus contrast produced results that were largely consistent with the predictions of the settling dynamics account. Additional methodogical and experimental procedures that were discovered in the course of this work and that may provide even stronger support for the account were also outlined.

Taken together, the first part of the thesis provided strong support for the settling dynamics account as a means of reconciling the complex and often contradictory effects of semantic ambiguity that have been reported in a number of studies. This work also undermined the claim that different configurations of the decision system were responsible for most of the different ambiguity effects that have been observed. However, this work demonstrated only the sufficiency of the settling dynamics account as a basis for explaining these effects and did not demonstrate that the activation in semantics, in fact, drives the response system.

To more directly evaluate whether semantic settling dynamics or the configuration of the decision system are responsible for producing different ambiguity effects, it would be useful to implement models of both of these systems (and potentially other systems, such as those responsible for orthographic and phonological processing) to simulate actual tasks. This would allow

for a more transparent evaluation of which dynamics are actually responsible for different effects. Moreover, such an integrated system would provide a means of assessing whether additional effects that were not captured by the settling dynamics account are an emergent property of how the decision system interacts with different representations. However, the development of such an integrated model is challenged by previous work that claimed that connectionist models are ill-suited for studying response selection.

To evaluate this claim, and the validity of the general principles that underly standard connectionist models, an extension of the original numerosity judgment experiment that produced data that were inconsistent with several connectionist models was completed. This study showed that several esoteric aspects of the original study were likely responsible for the discrepancies between the connectionist models and the behavioral data. Follow-up simulations further supported this inference and showed that a connectionist simulation that employed biologically-plausible processing constraints captured most of the effects in the behavioral data. Furthermore, these simulations demonstrated that standard error-driven learning principles were capable of capturing trial-by-trial sequential and adaptive effects that are outside the scope of the dominant models of decision making. Admittedly, some details remain to be addressed, but a careful consideration of the behavioral work and of the model's performance and underlying representation and processing principles have suggested several targeted improvements to future simulations. Thus, contrary to the original claims regarding the ill-suitedness of the connectionist framework for modeling response selection, a connectionist model is, in many critical respects, a superior model of response selection relative to the major task-specific accounts of performance in two-alternative forced-choice tasks (e.g., the diffusion model). Additionally, the outcome of this work highlighted the importance of considering how well a model of response selection can account for performance in a range of tasks, and in expanding the breadth of effects that are considered when developing theories of response selection.

Taken together, the outcome of the response selection work has indicated that essentially

the same learning, representation, and processing principles used in the biologically-plausible simulation of semantic ambiguity effects can provide a valuable framework for understanding performance in multiple domains. Moreover, this work suggests that such a framework could be used to simulate tasks that integrate, amongst other systems, how lexical-semantic knowledge can interact with the response system to generate a range of different ambiguity effects. The domain general nature of this framework, as well as the potential relevance of both semantic processing and response selection in driving particular patterns of effects, further suggests that developing simulations of a range of tasks that examine ambiguous word comprehension will lead to important insights into both systems.

In conclusion, the co-ordinated computational and behavioral investigations that were reported in this dissertation, albeit still in need of improvement with respect to some details, support the connectionist framework as a general model of a range of cognitive systems. They also provide a basis for developing more comprehensive and transparent accounts of a number of phenomena including the comprehension of ambiguous words and how the representations of these words drive response selection in different tasks. Additionally, the work highlights how these models, even in imperfect states, can serve to generate valuable predictions that guide further research in ways that are not possible in the context of narrow and underspecified theories.

An extension of the present work via follow-up simulations and behavioral investigations shows considerable promise for deepening the initial knowledge that has been generated in the course of this work. To this end, the present work will serve as a foundation for altering the breadth and scope of how accounts of phenomena in a range of domains are formulated and for asking and answering many questions that are often not even posed in the extant literature. This is undoubtably an extremely large undertaking and many unanticipated challenges will likely be encounted in the course of this endeavor. The rewards for success, however, indicate that such efforts are well justified, and the work completed as part of this dissertation provide a valuable theoretical and methodological framework for acheiving these goals.

Part IV

Appendixes

Appendix A

SOS! An Algorithm and Software for the Stochastic Optimization of Stimuli

All empirical researchers recognize that a fundamental challenge in the preparation of a new experiment is the selection of stimuli that are optimally suited to address the theoretical question of interest. The definition of what constitutes an optimal set can be unpacked into two main components: (1) the manipulation of the variable(s) of interest should be as large as possible in the absence of confounds with other variables (i.e., internal validity), and (2) the stimuli should be representative of their underlying population(s), thus permitting inferences to items not present in the experiment (i.e., external validity). Researchers also know that the consequences of failing to select an adequate set of stimuli can be dire. In the worst-case scenario, suboptimal stimuli can generate putative theoretical "advances" and protracted periods of belief in incorrect theoretical positions (Cutler, 1981; Gernsbacher, 1984). In a less extreme scenario, suboptimal stimuli can make it impractical to achieve the necessary statistical power to study the effects of interest (e.g., Armstrong, 2007). Even in the best-case scenario, the use of suboptimal stimuli wastes time and other resources in efforts to increase the statistical power of the experiment, such as by running

Note. A version of the work reported in this Appendix was also reported in Armstrong, Watson, and Plaut (2012).

additional participants.

Given these issues, a newly minted psychologist would undoubtedly expect a high level of rigor in the methods available for selecting stimuli, particularly given the progress that has been made in other aspects of experimental design and analysis. For example, in visual word recognition research, massive efforts have been devoted to identifying variables that influence performance (Cortese & Khanna, 2007), to creating databases and software that list the values of these variables for large sets of items (Brysbaert & New, 2009; Coltheart, 1981; Davis, 2005; Kučera & Francis, 1967), to increasing the precision of timing hardware and software (Brainard, 1997; Schneider, Eschman, & Zuccolotto, 2002), and to improving the statistical frameworks used to analyze data (Baayen et al., 2008; Clark, 1973; Raaijmakers et al., 1999). The state of the art with respect to the selection of experimental items would therefore come as a shock: the standard method for selecting stimuli is to do so by hand with the assistance of a sorting function in a spreadsheet application.

The aim of the present work is to bring methodological rigor to the stimulus selection process, with the ultimate goal of facilitating the development of experiments that are better suited for the empirical evaluation of researchers' hypotheses. To this end, this appendix serves to introduce SOS—an algorithm and software for the Stochastic Optimization of Stimuli used in experiments, based on a variant of a classic stochastic relaxation search (Kiefer & Wolfowitz, 1952). The appendix begins by characterizing the benefits and drawbacks of existing manual selection heuristics and computational search algorithms that can be used to identify optimal experimental stimuli. This is followed by a brief description of how SOS augments a simple manual optimization heuristic to rapidly and reliably discover optimized stimuli. This is accomplished by tailoring a general stochastic optimization algorithm so that it is well suited to the types of selection constraints and challenges to successful optimization faced by psychologists during item selection and other similar problems. Following the description of the algorithm, the results of Monte Carlo simulations of SOS performance in solving optimization problems that are encoun-

tered in a variety of experimental designs are reported. These examples serve to illustrate the robustness of the algorithm and its superiority to manual stimulus selection, the new types of experimental designs it facilitates, and the increased rigor it brings to evaluating and optimizing the internal and external validity of experimental items. These simulations also demonstrate that our software implementation of the algorithm can be used with relative ease to identify optimal stimuli. Note that although all of these example optimization problems involve matching word sets for use in psycholinguistic experiments—an area in which such optimizations may be particularly valuable—the algorithm is domain general and can be applied to any situation that involves selecting from a population of items characterized on a number of dimensions. Extended uses of SOS, such as for the selection of control participants for neurologically impaired patients in case series analyses, are briefly discussed in the final section. A user manual including additional details of the SOS procedure, example optimizations, and the full, open-source software implementation of the algorithm as either MATLAB source files or pre-compiled binaries for major operating systems is available free of charge at `http://sos.cnbc.cmu.edu`.

A.1 A manual Heuristic for Identifying Optimal Stimuli

The SOS algorithm has much in common with a simple manual heuristic for identifying optimal stimuli; indeed, the simplest version of the algorithm ("greedy" optimization of stimuli, discussed later) basically amounts to a formalization of this heuristic. It can be broken down into the following major steps:

1. Determine the conditions that will be employed to study the variable(s) of interest.[1]

2. Define constraints that establish how variables should differ or be equated across condi-

[1] For the purpose of illustration, well-known experimental designs are discussed to avoid unnecessary complexity. However, other superior but less familiar designs may be more suitable for testing some of these hypotheses (see, e.g., Baayen et al., 2008, for alternatives to standard multilevel designs in the study of continuous variables). How SOS can facilitate the selection of stimuli in some of these alternative designs is discussed in later sections.

tions or within a condition.

3. Identify a population of items and fill each condition with a sample of these items.
4. Search for an item in the sample and an item in the population that could be swapped to better satisfy the constraints.
5. Evaluate the degree to which the constraints have been satisfied (e.g., run *t*-tests to confirm that the variables that should differ between conditions do, in fact, differ). Repeat the search and evaluation steps until the constraints have been satisfied to a target threshold or until some other reason to stop the search has arisen, such as simply deciding to use the best set found to date after many unsuccessful attempts at swapping items.
6. Assess the degree to which the sample stimuli are representative of the underlying populations from which they originated. This step is often ignored but is necessary to draw statistical inferences to the broader population of items. Specifically, it is possible that the constraints have limited the items that are included in the sample to some contorted and unusually distributed subpopulation of the original population. Evaluating whether this is or is not the case is critical for determining the correct statistical analyses and inferences that can be drawn on the basis of the selected stimuli (Baayen et al., 2008; Clark, 1973; Hino & Lupker, 1996; Raaijmakers et al., 1999).

Problems with the manual heuristic

By and large, the manual heuristic just described is not unreasonable for selecting experimental stimuli. Nevertheless, it possesses several undesirable characteristics:

- The procedure is not actually optimal relative to automated optimization methods (support for this claim is provided in a later section and in van Casteren & Davis, 2007). This becomes increasingly evident as the complexity of the optimization problem increases. As was foreseen by Cutler (1981), the complexity of stimulus selection has only increased as time has passed, so these differences will likely be exacerbated in the future.

- The procedure is exceedingly tedious and time consuming and is, consequently responsible for the reuse of existing items. As a result, the generalizability of effects is hampered since the gold standard for generalization is replication with new items (Stanovich, 2001). The recycling of a particular set of stimuli can also lead to a single point of failure for a large body of work: if a problem is identified later with that particular set, the results of all the studies using those stimuli may be called into question. For instance, the ambiguous words used by Rodd et al. (2002), have been reused in studies by Armstrong and Plaut (2008) and Beretta et al. (2005). However, the results of these studies were called into question when later work by Armstrong and Plaut (2011) identified additional variables that were not controlled for in these items. Issues with the Snodgrass and Vanderwart (1980) pictures raised by Bunn, Tyler, and Moss (1998) also cast doubt on the outcomes of many studies employing these items.
- The procedure is rarely followed up with an examination of whether the selected stimuli are representative of the populations from which they were sampled. The lack of this step is conceivably due to a combination of factors, including the lack of a standard means of evaluating representativeness and the lack of pressure for these analyses to be included when results obtained using a particular set of items are reported. This oversight is troubling because the pressure to satisfy the constraints may distort how well the population is represented in the sample. Thus, the results of an experiment run with a particular set of stimuli may not generalize to the population from which the stimuli were sampled; the group of items that can satisfy the constraints may consist of only a small and possibly atypical portion of the population. Although this problem may exist for many different types of constraints a researcher could impose on the optimization process, factorial designs that cross intercorrelated variables may be particularly at risk of being nonrepresentative (see Baayen et al., 2008, for a discussion). The lack of an evaluation of representativeness is also somewhat paradoxical, given the long-standing debates concerning the use of

statistics to generalize the results of experiments both across items and across participants (Baayen et al., 2008; Clark, 1973; Hino & Lupker, 1996; Raaijmakers et al., 1999). These techniques are of little value if the population to which the results are being generalized is unknown.

- The procedure is prone to human error.

Given these problems, it is clear that the simple manual heuristic, although not without its merits, leaves much to be desired.

Alternatively, one avenue that intuitively might appear to address many of the problems with the manual heuristic outlined above would be to use a simple search algorithm from computer science to find optimal stimuli. However, these algorithms also tend to be unsuitable for selecting stimuli because exhaustive searches of all combinations of items are impractical and random searches of a subset of the possible combinations of items do not reliably identify optimal sets.

A.2 Stochastic Optimization of Stimuli: A Brief Overview

Considered as a whole, there are some advantages to both the human heuristic, which often yields stimuli that are sufficiently optimized for researchers to consider them suitable for use, and simple computational search algorithms, which allow very large numbers of combinations of stimuli to be examined but do not focus their efforts on exploring the combinations that are most likely to be optimal. A successful amalgamation of these two approaches might therefore leverage the unique advantages of both without the faults of either.

SOS is an attempt at such a synthesis. In its most basic form, the algorithm amounts to a translation of the steps from the manual stimulus selection heuristic into a "greedy" optimization search. In simple cases, this involves the following basic steps, although a range of more advanced options may also be applied. A detailed description of these steps, including their formal underpinnings, is presented in Subappendix 3 and is recommended reading for researchers who

wish to take full advantage of the algorithm.

1. Define the samples (experimental conditions).

2. Specify constraints that establish the desired relationships between the different variables within and/or between the samples, and associate each constraint with a *cost* function that operationalizes violations of the constraint. A broad vocabulary of constraints has been implemented in SOS that may be used to select items for a wide variety of experimental designs. These constraints are briefly described here, and the full details and formal underpinnings of the constraints are presented in Subappendix 1.

   Currently, the vocabulary of constraints can be divided into two main types: hard constraints and soft constraints. Hard constraints express all-or-none rules that restrict the values of a specific variable an item may have if it is to be included in a sample. For instance, *hard bound constraints* allow the user to impose specific upper and lower bounds on the values of a variable for all the items in the sample. These constraints take precedence over the soft constraints and serve to filter items that can potentially be included in a sample.

   In contrast, soft constraints do not operate on the basis of all-or-none rules; rather, they express all violations as matters of degree. For instance, a simple cost function for minimizing differences between two conditions *c1* and *c2* on a column of data containing each item's value on a variable, **x**, is to square the difference between the means of the two conditions on that variable:

$$O_{MIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2}) = (\bar{\mathbf{x}}_{c2} - \bar{\mathbf{x}}_{c1})^2 . \tag{A.1}$$

   Soft constraints are the main type of constraint that a user will employ during an optimization, and all have the same general form as the simple constraint outlined above. Many different soft constraints have been implemented to express a variety of desired relationships between the variables contained in one or more samples. These soft constraints can

be subdivided into two main classes: simple constraints and meta-constraints. Simple constraints allow for (1) minimizing or maximizing differences in the means or variances between variables at either the group or the item level between two samples or within a sample (*two-sample distance constraints*), (2) matching means or variances to target values (*one-sample distance constraints*), (3) distributing values of variables evenly across the range of the population or the current sample (*soft entropy constraints*), which may be particularly useful in mixed-effects/regression designs, and (4) matching the correlation between two variables to a target value (*correlation-matching constraints*), which is often zero so as to eliminate correlations between independent variables that increase collinearity and weaken statistical power.

Meta-constraints operate on the same principles as simple constraints but serve to constrain the values of other constraints. This effectively allows the user to control the importance and degree to which different constraints are satisfied, which may be relevant in some optimizations. These constraints can be used (5) to allow a constraint that maximizes the difference between two conditions to be optimized only once another constraint that minimizes the difference between two conditions has been minimized—for instance, to ensure that differences on nuisance variables are eliminated before maximizing the differences between conditions on a variable of interest (*conditional matching constraint*)—or (6) to require that two constraints be satisfied to the same degree (*matching constraint*).

The full set of cost terms that operationalize the constraints a user imposes on item selection constitutes the total cost of the set of stimuli—that is, the degree to which all of the constraints have been satisfied. For instance, in an investigation of word frequency effects using both a low- and a high-frequency word condition, the full set of constraints might consist of a two-sample distance constraint that maximizes the differences between the samples on word frequency, such that the mean word frequency is lower in the low-frequency condition than in the high-frequency condition, and other two-sample distance

constraints that minimize the mean groupwise differences on confounding variables such as word length.

3. Associate each sample with a population of items and fill that sample with a (random) sample from that population. This sample typically constitutes a very suboptimal sample, as expressed by a relatively high cost value (in comparison with other cost values obtained later in the optimization). Other initial sampling methods, such as specifying all or part of the items in a list, are discussed later in the appendix and in the user manual.

4. Attempt to improve the samples by randomly selecting two items, one from a particular sample and the other either from its population (*population feeder* item selection) and/or from another sample sharing the same population (*sample and population feeder* item selection), and swapping them. If the swap results in a reduction of the total cost, the optimization "moves" to the swap set; otherwise, the original samples are retained, and the optimization "stays" with the original set. Each of these attempted swaps is referred to as an *iteration* and can be accomplished in a computationally efficient manner via a *local update*. Local updates consist of calculating the difference in cost resulting from changing the two items, rather than re-calculating each cost term using the whole data set. For example, rather than recomputing the mean by averaging the sum of values across all items, the value of the item being swapped out can be removed from the previously calculated mean, and the value of the item being swapped in can be added to that mean (see the user manual for details).

5. Statistically evaluate how well the constraints have been satisfied—for instance, by using t-tests to determine whether differences across variables to be maximized have p-values less than .05 and differences that are to be minimized have p-values greater than .5. Other statistical tests are also available to evaluate whether a correlation between variables has been matched to a target value (often zero, to remove confounding covariates) or to evaluate whether the values of a variable are uniformly distributed (either across the range of values

in the sample or across those in the population, depending on the intended generalizations to be drawn from the items). If these statistical criteria have been met (or no improvement in the value of the cost function has occurred for a specified number of iterations, called the *freeze point*), the optimization terminates.

6. Assess the degree to which the selected items represent the broader population. This is accomplished by running the optimization procedure several times and comparing the degree of overlap in the selected items across different runs. High overlap suggests that only a restricted sub-population of items is capable of satisfying the constraints. Consequently, generalization of the results obtained with these items to the broader populations associated with each sample may not be appropriate. Conversely, low overlap suggests that results obtained with these items can be generalized to the broader population.

This greedy optimization routine is extremely rapid, by virtue of only making swaps that lower the overall cost of the samples, and may be sufficient to satisfy an experimenter's constraints in many cases. Additionally, the results of such an optimization are used to configure the more advanced stochastic optimization routine, as is discussed later. Consequently, a greedy optimization is prescribed as the first type of optimization to run. However, greedy optimization has a drawback in that making only swaps that decrease the cost may sometimes result in the algorithm becoming prematurely stuck in a *local minimum*, where swapping any one pair of items would result in a cost increase. Thus, despite the fact that a better *global minimum* cost value may exist elsewhere in the set, reaching it is not possible without at least some swaps that increase cost. The existence and likelihood of becoming stuck in such a local minimum is a function of several characteristics of a given optimization problem and is usually not explicitly determined, because this would require the use of a computationally impractical exhaustive search of all possible samples. As a general guideline, however, the likelihood of becoming trapped in local minima increases as sample size increases, as population size decreases, and as the number and complexity of the constraints increase (see Hinton & Sejnowski, 1986, for additional discussion).

Stochastic optimization

A principled means of avoiding these local minima is instantiated in the stochastic variation of the optimization search. Here, instead of applying a strict "move to the set of items with the lowest cost" rule, the cost difference, $\Delta cost$, between the original set of items and the swap set only biases the likelihood of choosing the set with the lower cost, with an additional term adding random variability to this choice as in classic stochastic relaxation searches (Hinton & Sejnowski, 1986; Kiefer & Wolfowitz, 1952; Rumelhart, Smolensky, McClelland, & Hinton, 1986). The variability in the decision to swap to a different set of items formally guarantees that an optimal set will be found if the optimization is run for a sufficient number of iterations, since it allows the algorithm to escape from local minima.

The amount of random variability can be manipulated via a parameter referred to as *temperature* (T), where higher temperatures increase the random element of the decision and where temperature has a lower bound of zero. The likelihood of swapping a pair of items is determined in a sigmoid function that depends on both the $\Delta cost$ and temperature parameters, as formalized in the following equation:

$$p_{swap}(\Delta cost, T) = \frac{1}{1 + e^{\frac{-\Delta cost}{T}}}.$$

Figure A.1 illustrates this relationship for several sigmoid functions generated with different values of temperature and $\Delta cost$. As this figure shows, temperature modulates the degree to which $\Delta cost$ can bias the likelihood that a swap that reduces cost will take place: an infinitely high temperature causes swaps to be made without any consideration of whether the swap increases or decreases cost, whereas a temperature value of zero eliminates random variability and is equivalent to a greedy search. Consequently, fully leveraging the stochastic algorithm requires using an intermediate temperature value that ultimately leads to reductions in cost without causing the algorithm to become stuck in local minima.

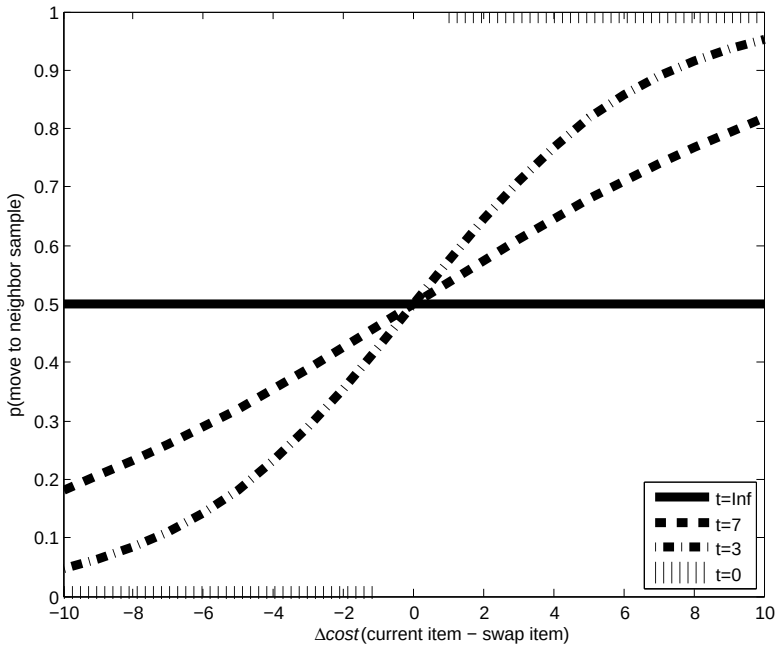


Figure A.1: Depiction of the likelihood of "moving" to the "swap" item or "staying" with the current item based on a range of values of the cost difference, $\Delta cost$, between the two items that could be swapped, and temperature. Inf = infinite.

Rather than attempting to identify a single temperature value, preliminary simulations have found, as others have before, that more rapid and reliable optimizations are possible by using an *annealing* function that gradually lowers temperature throughout the course of the optimization. Specfically, temperature decays at an exponential rate that has been shown to be optimal in related optimization problems (Ackley et al., 1985; Kirkpatrick, Gelatt, & Vecchi, 1983). The detailed overview in Subappendix 3 and the example presented in Subappendix 2 describe how this temperature annealing schedule is calibrated on the basis of the results of the greedy optimization. Essentially, this process simply involves initially setting temperature to a sufficiently high value such that the vast majority of swaps are being biased by $\Delta cost$ only to a trivial degree. Subsequently, temperature is gradually lowered toward a $\Delta cost$ value that allows for a small percentage of swaps to still occur when an equivalent greedy optimization reaches its freeze point. This allows for the algorithm to escape from the minimum found in the greedy search if it is only a local one.

A.3 SOS Performance on a Range of Optimization Problems

The next section reports the results of Monte Carlo simulations on several example optimization problems to demonstrate the value and robustness of SOS. In the first of these examples, The first portion of this section examines a set of previously published psycholinguistic experiments for which the stimuli were manually selected to conform to a standard one-way design that dichotomizes a continuous variable (Morrison & Ellis, 1995). Using the same population of items these researchers had at their disposal, the results of manual selection are compared to those produced by SOS and show that SOS yields superior matches. This is followed by a report of how SOS can be used to construct another set of stimuli optimally suited for use in what is, arguably, a superior mixed-effect regression analogue of the original one-way designs used in two separate experiments.

These detailed reports are followed by brief reports the results of several additional realistic

optimizations that correspond to experimental designs often encountered by researchers. Again, note that although all of the examples concern psycholinguistic research questions, the experimental designs that are covered are representative of the designs that psychological researcher might encounter, irrespective of his or her domain of inquiry. The scripts that contain the commands for each realistic case are available via the online user manual and may be useful starting points for users who want to create their own optimizations.

Comparison of SOS stimuli with manually-selected stimuli in a two-level, one-way design

As a first illustration of the SOS algorithm, the performance of SOS-selected items is compared to the items selected manually in an influential study by Morrison and Ellis (1995, , hereafter, ME95). They examined whether the classic "word frequency" effects observed in a range of tasks (e.g., Schilling, Rayner, & Chumbley, 1998) have been overestimated because of the correlation between word frequency and another variable, the age at which words are learned, or "age of acquisition" (AoA; Gilhooly & Logie, 1982). To try to disentangle these effects, ME95 designed several experiments in which frequency and AoA were each varied independently while the other variable was held constant. As an example case for SOS, we will examine the stimuli ME95 designed to examine the effects of word frequency in isolation (Experiment 2): 24 pairs of high- and low-frequency words that differed on frequency while being matched on AoA, word length (in number of letters, expressed in the *letters* variable), and imageability.

On the basis of ME95's description of the way in which they selected stimuli, the population of items they had at their disposal was recreated by taking the union of the items having both AoA and imageability ratings (Gilhooly & Logie, 1980a) and word frequency data (Kučera & Francis, 1967) in the MRC psycholinguistics database (Coltheart, 1981). In this population, the correlation between AoA and frequency was .22 and ranged from .04 to .67 between each of these two variables and the length and imageability variables. Next, following ME95, high-frequency population was created that included words with frequencies greater than 110 per million, and a

low-frequency population was created that included words that occurred 10 times per million or less. By manually selecting 24 pairs of stimuli from these two populations, ME95 were able to vary the high- and low-frequency stimulus lists significantly on frequency ($p = 2.5 \times 10^{-8}$ based on a two-tailed t-test). The two lists were matched perfectly on length ($p = 1.0$). However, the two lists were matched less well on AoA ($p = .20$) and not at all on imageability ($p = .01$).

With the aim of outperforming manual stimulus matching, SOS was used to select high- and low-frequency words from the same population as that used by ME95. A tutorial detailing the full process of running this optimization in the SOS software is included in Subappendix 2. In brief, this process involved creating constraints indicating that the words in the low-frequency condition should have lower frequencies than the words in the high-frequency conditions, constraints indicating that the words in both conditions should be matched on length, AoA, and imageability, and additional constraints that allowed for frequency differences to be maximized only once length, AoA, and imageability had been relatively well matched (see the subappendices for further details that motivated these constraints). The algorithm was then configured to continue to run until the two lists differed on frequency ($p < .05$) and were matched on AoA, length, and imageability, $p > .5$. All tests and constraints were pairwise in nature, as in ME95.

Using an initial greedy optimization, the parameters for the temperature annealing function used in the stochastic optimization were calibrated on the basis of the assumption that the greedy version had become prematurely stuck in a local minimum (as described above and in detail in Subappendix 2). A stochastic version of the algorithm (refered to as the SOS optimization) was then run. This optimization found a solution that exceeded the statistical criteria that were set: the high- and low-frequency samples that were selected varied on frequency [paired $t(23)$ = 16.13, $p = 4.92 \times 10^{-14}$], but not on AoA [paired $t(23)$ = -0.68, $p = .51$], length [paired $t(23)$ = 0.33, $p = .75$], or imageability [paired $t(23)$ = 0.39, $p = .70$]. These results show that SOS quickly found sets of stimuli that are better suited to test the theoretical question posed by ME95. In particular, the ME95 samples were less significantly different in terms of word frequency, while

also being less well matched on AoA and imageability (which differed significantly across their two samples).

The descriptive statistics for the high- and low-frequency samples identified by ME95 and two SOS optimizations are presented in Table A.1 and offer additional insight into how SOS obtained better matches (the SOS[length] optimization is described in detail later). This table shows that whereas both SOS optimizations generally found approximately equivalent or substantially better matches than did ME95 on AoA, length, and imageability, the mean difference in word frequency was, in fact, smaller in these samples than in ME95. Nevertheless, a more significant effect of word frequency resulted, because there was increased variability in the difference scores for the ME95 match; the effect size of the frequency difference was 1.7 for the ME95 matches versus 3.3 for the SOS optimization. This should translate into decreased within-condition error when behavioral data related to these items are analyzed and a reduced reliance on the data from the few items with the most extreme frequencies driving up the mean difference in ME95. Without going into detail here (but see Subappendix 1), the cost function used by the pairwise minimization constraints also served to minimize any correlation between the two samples in ways that are not reflected in the t-tests but do manifest themselves in measures of collinearity, such as the variance inflation factor. For instance, the variance inflation factor for the frequency effect was 16% higher in the ME95 matches than in the SOS optimization, which would lead to more powerful frequency inferences if the matched variables were included as covariates. Taken together, these results serve to demonstrate how the constraints instantiated by SOS can lead to the discovery of alternative, superior item sets, sometimes in potentially unexpected ways.

An examination of Table 10.1 reveals a further characteristic of the optimization that is worth additional discussion: the high- and low-frequency samples discovered in the SOS optimization, although nonsignificantly different on length ($p = .51$), are not as well matched on the variable of length as those reported by ME95 ($p = 1.0$). This outcome is a natural effect of the SOS algorithm: all constraints will tend to be matched to the same level unless a particular constraint

is prioritized. Insofar as there is no strong theoretical basis for prioritizing a particular constraint, such behavior is usually desirable, since it tends to eliminate all confounds to approximately the same extent. However, if precise matches on one variable are particularly important for theoretical reasons, SOS can be configured to attempt to satisfy the constraints imposed on this variable to an increased degree. To illustrate this capacity, an additional optimization was run in which a weight of 100 was added on the length constraint. The statistical criterion for the length variable was also changed to require a perfect match between the conditions ($p = 1.0$). This weighting parameter was chosen so that the initial cost values associated with length dominated overall cost. The choice of a value of 100 is not specific to this problem and generally produces this type of distribution across the cost terms, because SOS operates on the normalized values of each variable, as described in Subappendix 3. Additionally, the statistical criterion for the frequency test was modified such that an equally significant difference in frequency was required to stop the optimization as was obtained in the human match ($p = 2.5 \times 10^{-8}$). This change was made because, otherwise, the more lenient p-value criterion of .05 was exceeded, and the optimization was stopped before much better sets were discovered. In this new optimization, the high- and low-frequency samples were perfectly matched on length ($p = 1.0$), while still exceeding the original statistical criteria that were stipulated and the matches obtained by ME95 (AoA, $p = .81$; imageability, $p = .82$; frequency, $p = 3.2 \times 10^{-14}$). For comparison purposes, Table 10.1 also lists the descriptive statistics for this additional set of items (the SOS[length] optimization).

In addition to evaluating the internal validity of the items produced by SOS during the course of the optimization, the degree to which this optimization selected stimuli that were representative of the underlying population was also assessed. As was described earlier, it is not possible to make claims about the generalizability of any experimental effects observed with these stimuli if they represent an idiosyncratic subset of the population. To evaluate the generalizability of the SOS stimuli, five different stochastic optimizations were run until each passed the statistical

criteria . The overlap in these samples was then calculated to determine how similar the resulting lists in each condition were to each other. On average, the high-frequency lists shared 13% (SD = 0.09) of the total number of items across each optimization, while the low-frequency lists shared 8% (SD = 0.06; values of 25% and 8% were obtained in analogous comparisons for the items discovered in the SOS[length] optimization). These values suggest that the five optimal solutions were, in fact, quite different from one another and that the algorithm was not discovering the same unique solution each time. Thus, any behavioral effects observed with these stimuli should generalize to new sets of stimuli on the basis of the principles of statistical inference. Specifically, many different items can be chosen to satisfy the specified constraints. Using any one of these samples thus corresponds to using one of many possible random samples of a broad population of items. Consequently, the results obtained with any one of these samples are likely to generalize to the population. Nevertheless, it is worth noting that generalization based on statistical inference alone can occasionally be misleading if the random sample that is selected is unusual in an idiosyncratic manner purely by chance. The most confident generalizations would thus also be supported by running more than one experiment, using different random samples of items that satisfy the experimental constraints. By evaluating the overlap between several sets of items generated by SOS, sets of items suitable for this task are already available for this purpose.

A regression analogue and expansion of the one-way design used by Morrison and Ellis (1995)

In the original ME95 work, the authors examined the effects of word frequency while holding AoA constant in one experiment (Experiment 2) and of AoA while holding word frequency constant in another experiment (Experiment 1). These inferences could also have been made in a single experiment using a standard 2×2 factorial design that crosses frequency and AoA. Such a design would carry the benefit of ensuring that the significance of a given effect was ascertained in the context of the same levels of the other effect. This is because the same mean AoA and

frequency values would be used in all of the statistical comparisons and this need not be the case across different independent experiments, which could confound the results. Practically, however, using such a design is not possible on the basis of how the conditions were defined in ME95; the hard cutoffs for inclusion in the high- and low-frequency and early- and late-AoA conditions used in Experiments 1 and 2 do not leave enough items in the four cells of the proposed design to create samples of the same size as in the ME95 experiments, even before stimulus optimization. This also suggests that somewhat different populations of items were used to study each type of effect.

The inability to simultaneously study the effects of AoA and word frequency may, however, be a result of the restrictiveness of a factorial design. Specifically, it is often the case that researchers will impose hard bounds on the items that can enter into particular cells of the design so as to maximize the differences observed between conditions. Presumably, this is part of what motivated ME95 to employ the low-frequency upper bound of 10 and the high-frequency lower bound of 110 in their second experiment and to employ similar constraints on the early- and late-AoA samples in their first experiment. This strategy is problematic for several reasons. First, sampling items in this manner allows only items from the extreme ends of the continuous variables under study to be selected for use in the experiment. Interpolation to this range using a standard linearity assumption, as in many analyses in psychology is questionable in this case because there are no data with which to evaluate this assumption. Second, finding items to fill the "atypical" cells in the design (e.g., the high-frequency, later-acquired words, which are rare given the negative correlation between frequency and AoA) is often difficult—if not outright impossible—because of the relatively small populations of items that can be considered for inclusion in these cells. Third, these designs can often be underpowered because within-condition variance on the variable of interest is transferred into the error term of standard statistical analyses. For example, words with frequencies of 110 and 11000 would both be grouped together in the high-frequency condition, and this more precise frequency information would not enter into

the analyses.

An alternative design which does not suffer from these problems is to sample a broad range of items that can be used to study the target effects via mixed-effects/regression analyses (e.g., Baayen et al., 2008). To do so, SOS wsa used SOS to select 100 items with a uniform distribution over the range of AoA and word frequency values in a slightly restricted version of the population used in the first example.[2] This was accomplished using entropy constraints. Constraints were also added to eliminate the correlations between the two main variables of interest, as well as between each main variable, length, and imageability, so as to minimize collinearity and boost the power of the regression. Statistical criteria were selected such that the optimization would terminate when all of the correlations were nonsignificantly different from zero ($p > .5$) and when the distributions were nonsignificantly different from uniform distributions ($p = .001$; Subappendix 1 discusses the motivation for using a smaller p-value when the uniformity of a distribution).

Initial runs of SOS showed that the statistical tests related to the correlations were being passed but that the optimization terminated by reaching the freezing point when the frequency and AoA distributions still differed from uniform distributions (e.g., frequency was still positively skewed with few high frequency items in the sample). Alternatively, if only the distribution constraints and statistical tests related to frequency and AoA were included, the algorithm successfully found a satisfactory solution (albeit one with significant intercorrelations among the different variables). Which of these solutions is most desirable is debatable, because each reflects a different trade-off of internal and external validity and statistical power. Moreover, these two results suggest, as would be expected on the basis of the original ME95 experiment structure,

[2]SOS was unable to create a uniform distribution over the full range of the frequency variable in the initial population of items, even when this was the only constraint in the optimization. This was because the frequency data in the population were strongly right-skewed and there were insufficient high-frequency items to sample the upper end of the frequency continuum uniformly. To avoid this problem, the top 5% of words with the highest frequencies were trimmed.

that pulling apart these two dimensions is a nontrivial task. Nevertheless, an attempt was made to find a compromise between these two extremes by adding a weight to each of the frequency and AoA distribution constraints so that they were of the same or a larger order of magnitude as those of the correlation constraints at the beginning of the optimization; a weight of 100 satisfied this goal.

An initial Monte Carlo simulation showed that the updated version of the greedy optimization succeeded in satisfying the statistical criteria 8/10 times. Using the results of a randomly selected failed optimization, a stochastic version of this same simulation was configured. In this case, the optimization succeeded 10/10 times. To illustrate the effect of the entropy constraints on creating uniform distributions, the initial and final AoA and frequency distributions are presented in Figure A.2.

Additionally, the overlap beteween the 10 samples that were discovered by the stochastic optimization was computed. The overlap between these samples was 42%, suggesting that the satisfaction of this constraint does depend in part on a subset of these items. Consequently, a researcher might wish to consider relaxing some of the statistical criteria if the external validity of the experiment is paramount.

Performance of SOS on additional examples

Next, the results of several realistic optimizations corresponding to additional experimental designs often encountered by researchers are briefly reported. These optimizations demonstrate the capabilities of SOS in a variety of other real-world scenarios. The examples are based primarly on the word data from the MRC Psycholinguistic Database (Coltheart, 1981) that have values for Kučera and Francis (1967) frequency and the number of letters, phonemes, and syllables for each word, unless otherwise specified. Variables were considered matched if $p > .5$ and significantly different if $p < 0.05$. All examples below were run as stochastic optimizations unless otherwise stated, and the number of successful optimizations (out of 10) is also included. The script files

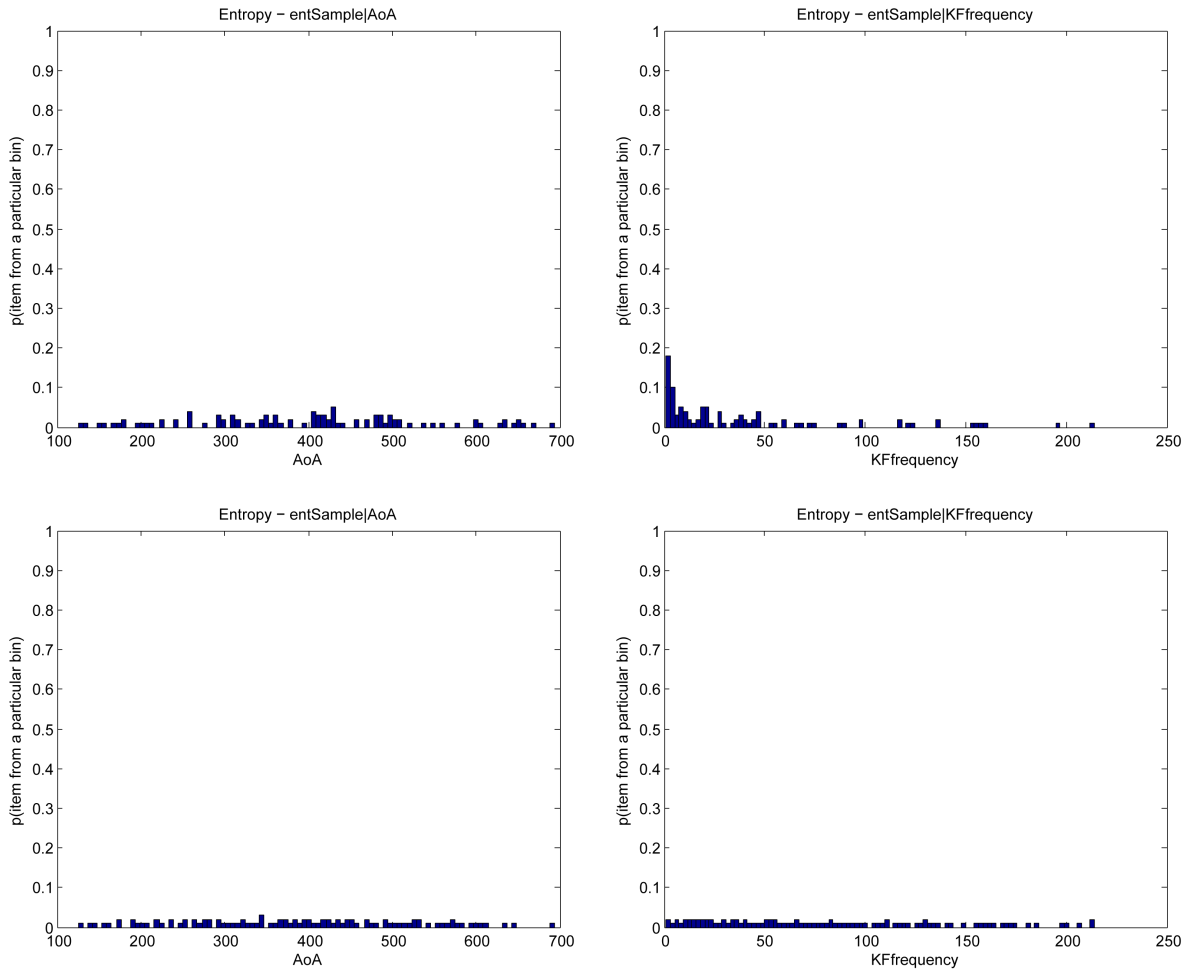


Figure A.2: Plots generated by the SOS software of the pre- and post-optimization AoA and frequency distributions when a constraint was imposed to pressure the creation of uniform distributions across the populations (AoA range: 125-694; frequency range: 1-214). The pre-optimization AoA and frequency distributions had entropy values of -0.012 and -0.071, and these entropy values increased to -0.0095 and -0.0061 by the end of the optimization. These final distributions did not differ from a uniform distribution ($p > .001$).

for these and additional designs are available in the user manual.

Categorical/ANOVA designs.

One-way designs. Several one-way designs were that involved identifying two samples with 100 words in each sample. These designs are similar to the ME95 case described above (e.g., matching two lists for length, number of syllables, and number of phonemes, while maximizing frequency differences) but did not involve the stipulation of upper and lower bounds for inclusion in the two different samples. Instead, SOS discovered the optimal separation between these distributions while sampling all items from the same population. These optimizations succeeded 10/10 times when matched groupwise.

A similar optimization also succeeded 10/10 times when items were matched pairwise, with the following caveat. Because of the details of the cost function used to minimize pairwise differences between conditions, pairwise matching also reduces the variance in the pairwise differences. This sometimes produces samples that, while differing only very slightly in mean values, yield p-values less than .05 for variables intended to be matched (see Subappendix 1 for additional discussion). In one instance of this optimization, for example, the mean number of phonemes in the high-frequency sample was 5.08 (SE = 0.17), while the mean number of phonemes in the low-frequency sample was 5.02 (SE = 0.18). Although a difference of 0.06 phonemes is very unlikely to affect the behavior of participants in a substantial manner, the low variance in each list caused the samples to be under the $p > 0.5$ cutoff for matching constraints, $t[paired](99) = 1.23, p = 0.22$. To prevent these effectively "good" samples from being rejected by SOS, a "threshold" parameter was implemented that allows statistical tests in SOS to be passed if the mean difference between two samples is less than a specified amount. In the example discussed here—two samples maximized for frequency differences but matched pairwise on length, phonemes, and syllables—we dictated that optimizations could pass if the variables to be matched had means within 0.5 of each other. Depending on the problem at hand and the

scale of the variables, users may opt to not use thresholding at all or to change the cutoff to more appropriate values for the particular samples they are attempting to create.

The performance of SOS was also examined when three, rather than two, final samples were desired; in other words, the experimental design in this case corresponds to a one-way design with three levels of frequency (low, medium, and high). Additionally, all of these conditions were matched for length, number of syllables, and number of phonemes. The three levels of frequency were not stipulated in advance. Rather, SOS discovered an optimal parcellation along the frequency variable, in which each level differed significantly from the others, and the intermediate level was approximately equidistant between the low and high levels. This optimization produced acceptable stimuli 10/10 times matching items group-wise and 10/10 times matching items pairwise (a threshold of 0.1 was used when the samples were matched pairwise).

Two-way designs. Often, researchers want to examine the influence of two variables simultaneously. These 2×2 designs often pose substantial difficulties for manual selection heuristics, especially when each cell in the 2×2 design also needs to be matched for multiple confounding variables (as in the ME95 example). Nevertheless, provided that sufficient items exist in the population to create such a design, SOS is capable of discovering appropriate samples. To demonstrate this, SOS was used to create a 2×2 design that crossed the effect of word frequency with that of word length, while minimizing any differences on imageability across these four samples. Items were drawn from a single population containing all of the words in MRC for which length and AoA data were available and for which word frequencies were constrained to be less than or equal to 1,000. This last restriction was imposed for the same reasons as described in the regression analogue to the ME95 design. No restriction on which type of item could be placed into a high or low cell on a particular variable was imposed such that the optimizer was required to discover an optimal division between the conditions. SOS found samples that satisfied these constraints 10/10 times with both groupwise and pairwise constraints, although threshold parameters were required in the pairwise version of the optimization.

Other optimization problems. Another feature of SOS is the ability to specify some stimuli as being "fixed." This allows SOS to discover a sample that possesses certain relations to a predetermined list—for instance, items from a previous experiment that is being extended. For example, we asked SOS to find a sample of 100 nouns that were matched to 100 preselected verbs for frequency, length, and number of syllables and phonemes. This optimization was successful 10/10 times when matched groupwise and 10/10 when matched pairwise (a threshold of 0.1 was used when the samples were matched pairwise). In a similar vein, SOS can also be used to find two samples of words matched to specific, user-inputted values of frequency while still matching these samples on length, number of syllables, and number of phonemes. This optimization was successful 10/10 times group-wise and 10/10 times pairwise.

Additional types of optimizations, including but not limited to matching the variability across conditions, matching to a target correlation, spreading scores uniformly within a sample rather than across the range of the population, and replacing only a subset of items in a sample, are also possible in SOS (see Subappendix 1 and the user manual for details).

A.4 Extended Applications

Although the discussion to date has focused primarily on the selection of stimuli for new psycholinguistic experiments, SOS can be applied to a wide range of optimization problems. Moving beyond stimulus selection, one exciting potential application of SOS is in *participant* selection, particularly in the context of investigating the effects of a neurological dysfunction such as semantic dementia, herpes simplex encephalitis, or stroke on a cognitive system such as semantic memory. Patterson and Plaut (2009) have recently argued that case-series analyses (e.g., Woollams, Lambon Ralph, Plaut, & Patterson, 2007) may be more useful than single-case studies when studying the mechanisms that underlie these systems and their neural substrates. Conducting such case-series studies, however, is still quite a challenging endeavor because of the need to match individual patients with healthy controls on a number of variables (e.g., age, years of

education, performance on baseline intelligence tests). Using SOS to select these controls could greatly enhance a study's ability to inform researchers on the underpinnings of a particular type of impairment.

Another application of SOS concerns the re-analysis of a subset of items from an experiment that was found, post hoc, to use a suboptimal set of items. For instance, Watson (2009) has matched two categories of verbs, those with strong or weak associations with specific motor programs, on published imageability ratings. Later, she discovered that the published verb ratings did not correlate with ratings elicited by participants when verbs were rated within their grammatical context (e.g., "the race" versus "to race"). Moreover, a regression-based solution to this problem was suspect because the ratings were bimodally distributed. By selecting an optimal subsample of items with SOS (treating the existing sample items as the population), Watson was able to reanalyze these data without the presence of any imageability confounds. SOS was also successfully used for a similar problem where population and/or dialect differences led to a confound on familiarity measures in the items used by Armstrong and Plaut (2008) and in selecting a larger set of items for use in follow-up work (Armstrong & Plaut, 2011).

Another use of SOS has been to leverage data from behavioral piloting for functional magnetic resonance imaging (fMRI) experiments to preemptively select ideal stimulus sets. With standard block-design fMRI experiments, it is often desirable to have different conditions (blocks) equated for accuracy and reaction time, ensuring that differences in neural activity are not due to difficulty confounds (Binder, Desai, Graves, & Conant, 2009). Frequently, researchers collect behavioral data on the tasks that will be presented in the scanner, and, if accuracy or reaction time differences exist, they are forced either to manually select subsets of blocks or individual trials that are equated on these measures, or to include these measures as covariates in analyses of the BOLD data itself. When Watson and Chatterjee (2012) encountered this issue, they instead used SOS to select items from the behavioral pilot experiment that were equated for difficulty across conditions.

A.5 Using and Modifying SOS

One guiding principle in the development of the SOS algorithm and software was that it would be impossible to anticipate all of the potential uses and insights other researchers might have with regards to the optimization procedure. The SOS software was therefore written in a standard, multiplatform, easy-to-program environment (MATLAB) that is already familiar to many psychologists. The source code for the software has been made freely available for academic purposes. The use of object-oriented programming practices, combined with the implemented examples of different constraints and cost functions discussed in this appendix, the user manual, and the source code documentation, should also facilitate the modification and extension of this code for new purposes.

A.6 Conclusion

Despite the critical importance of selecting optimal stimuli, this aspect of experimental design has largely been ignored, while other aspects have benefited from massive advances. The present work outlined some of the obvious and not-so-obvious effects that suboptimal stimuli can have on the internal and external validity of experiments and, ultimately, on the overall value of a particular study. It also provided a theoretical basis and empirical support for a solution to these problems—the SOS algorithm and software—which implements an optimization tailored for easy use in psychological research. This tool allows researchers to focus on choosing an optimization problem that suits their theoretical question, rather than on the tedious task of manually selecting stimuli and the inherent limitations of that method. It is hoped that many researchers will find this tool useful when selecting experimental stimuli, and, looking ahead, that SOS will be further developed and extended to enhance many other aspects of experimental and analytical procedures.

A.7 Subappendix 1: Implemented Cost Functions

A number of constraints and associated cost functions have currently been implemented in the SOS software and can broadly be divided into two categories: hard constraints and soft constraints. Hard constraints effectively amount to strict inclusion/exclusion rules governing the values of a variable in a condition and are functionally equivalent to filtering out items that exceed specified cut-offs from the population before beginning to optimize the stimuli. For instance, a hard constraint can be used to specify that all of the word frequencies in a low-frequency condition must be below a particular value. In contrast, soft constraints have continuous cost functions that denote the degree to which a set of stimuli satisfies the constraints. In principle, all possible sets of stimuli could be selected to satisfy the constraints to greater or lesser extents.

Because the interesting and practically useful aspects of the stochastic optimization algorithm relate to maximally satisfying soft constraints—particularly of the "simple" type, introduced in the next section—the main focus of this work has been to enrich the vocabulary of these soft constraints for users to employ in their own optimizations. Consequently, the soft constraints first and the issue of hard constraints is retured to later.

Soft constraints

Simple soft constraints. Simple soft constraints are soft constraints that directly reflect differences in normalized values on specified variables. These are the main type of constraint that needs to be stipulated before beginning the optimization. There are four main types of simple soft constraints: single-sample distance constraints, two-sample distance constraints, entropy (uniformity) constraints, and correlation-matching constraints. Two-sample soft distance constraints are introduced first because they are likely the most widely applicable type of constraint.

Two-sample distance constraints. This first type of constraint is concerned with minimizing or maximizing the distance between two samples of data on a particular measure, $f(\mathbf{x})$, at

either the group or the item level. Currently, this distance measure corresponds to either the mean or the standard deviation of each sample on a particular variable. In most applications, the data are expected to correspond to the values of the same variable across two different conditions (e.g., maximize the difference in word frequency across a low- and a high-frequency word condition). This notwithstanding, the software implementation also allows for the data from two different variables to be compared within a single condition or for two different variables to be compared across different conditions. For simplicity of notation, it is assumed that the cost function is measuring the same variable across two different conditions in the following discussion.

As was mentioned previously, maximizing or minimizing the distance between two samples on a particular measure applied to the item data for each sample, $f(\mathbf{x})$, can be operationalized using the simple cost penalty function $O_{MIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2})$ in Equation A.1. In past experience using SOS, this function has proven to be a reasonable starting point for operationalizing the desired relationship between the two variables on $f(\mathbf{x})$. However, in some instances, it may be desirable to prioritize the importance of some constraints over others. To this end, a weighting coefficient, b (with a default value of 1.0), and an exponent, n (with a default value of 2.0), have been added as free parameters. To accommodate the usage of exponents, the absolute value of the difference between the two measurements must now also be taken so that all differences produce a positive value. The result is a more general minimization function for group-level differences:

$$O_{MIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2}, b, n) = b\left(\left|f\left(\mathbf{x}_{c2}\right) - f\left(\mathbf{x}_{c1}\right)\right|\right)^{n}. \tag{A.2}$$

A maximization function for group-level differences can be generated in a similar manner. A particular ordering of conditions can be enforced by adding the additional sign function $s(f(\mathbf{x}_{c1}), f(\mathbf{x}_{c2}))$ that equals 1 when *c1* is less than *c2* and -1 otherwise:

$$O_{orderMAX}(\mathbf{x}_{c1}, \mathbf{x}_{c2}, b, n) = -s\left(f\left(\mathbf{x}_{c1}\right), f\left(\mathbf{x}_{c2}\right)\right)\ O_{MIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2}, b, n). \tag{A.3}$$

Manipulating either the weight or the exponent allows for the prioritization of each cost function relative to the total cost. As a general guideline, small manipulations of the weight allow for

more subtle emphasis of some constraints over others, whereas manipulations of the exponent cause more heavy-handed changes in cost. Weight manipulations are reminiscent of the β coefficients in a linear regression. This allows a researcher to prioritize the importance of satisfying constraints imposed on different variables on the basis of existing regression analyses that denote the degree to which each variable influences performance.[3] Exponent manipulations can be used to quickly increase or decrease cost to well above or below the level of other constraints. Such a manipulation effectively causes these constraints to be satisfied first before satisfying the other constraints. The details of how this occurs are discussed in the section on temperature annealing in the main text and in the user manual.

In addition to satisfying group-level constraints, similar constraints can also be imposed at the item level, provided there are equal numbers of items in the two conditions being compared. These pairwise constraints provide a tighter restriction on the relationships that exist across the items in the conditions being compared at the expense of being harder to satisfy. Pairwise constraints may be particularly beneficial in two situations. The first is, unsurprisingly, when a researcher desires to run pairwise statistical tests across items, thereby increasing the power of the experiment. The second situation occurs when, after matching a variable across conditions, any remaining variance due to the variable is to be extracted from the dependent variable using a regression-based analysis (e.g., multiple/mixed-effects regression, ANCOVA). In this case, controlling for a group-level measurement of the variable, such as the mean, may still lead to different distributions of that variable across the conditions. For example, imagine trying to match on word length at a group-wise level when creating high- and low-frequency conditions in a factorial design. Even by successfully matching the mean word length across the two conditions, it is still possible that the distribution of word lengths differs between them. For instance, word

[3]The relationship between β in regression and b in the cost functions is similar, but nonidentical, due to the nonlinearities in the cost and swap likelihood functions, although the two behave similarly. The similarity of this relationship is particularly strong in the context of multiple minimization constraints all being satisfied to similar degrees and when $\Delta cost$ falls within the relatively linear range of the sigmoidal p(swap) function.

length might be skewed positively in the low-frequency condition and negatively in the high-frequency condition. When these two distributions are merged into a single "length" variable and are regressed out, the composite length distribution may be correlated with the low- and high-frequency conditions.[4] Such a situation can lead to an interaction among the independent variables and a reduction in the variance in the dependent measure that can be uniquely ascribed to the effects of either length or frequency in the statistical analyses. Relatedly, this may enlarge the estimates of the variance for each predictor (i.e., increase the variance inflation factor). In contrast, if the distributions across the two conditions were matched on word length, as would occur if pairwise matching across stimuli is employed, such an interaction would not be possible. This can increase the amount of unique variance that is ascribed to the effect of word frequency and boost the statistical power of that effect. This is demonstrated in the regression-analogue example case in the main text.

Pairwise constraints require only a simple extension of the groupwise constraints presented earlier. The only difference between the two is that instead of calculating the particular measure of interest (e.g., mean) once across each condition before subtracting the measure of one condition from the other, the values for each *pair* of items, i, are subtracted from one another before calculating the measure. The absolute value of measures for each of the pairs can then be summed across all pairs, k, and the average measure can be extracted by dividing this sum by the number of pairs. For instance, to minimize the difference across all k pairs of items in two different conditions, the following cost function is used:

$$O_{pairMIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2}, b, n) = \frac{b}{k}\left(\sum_{i=1}^{k} \left|f\left(\mathbf{x}_{c2,i} - \mathbf{x}_{c1,i}\right)\right|\right)^{n}. \tag{A.4}$$

The same basic adjustments can be applied to the cost function used to maximize group-level

[4]Note that such a relationship may be present even if a pairwise t-test across the values in each condition produces a p-value of 1.0. This is because the t-test assesses whether the mean deviation across pairs—as opposed to the sum of the absolute values of the deviations across pairs—equals zero. A t-test p-value of 1.0 is therefore not sufficient to guarantee that this problem has been avoided.

differences, yielding a pairwise maximization cost function.

Single-sample distance constraints. A simplified variant of the two-sample soft distance constraint has also been developed that allows differences on a particular measure to be assessed relative to a user-specified value, v, instead of against the value of that measure in another sample. This type of constraint may be useful when generating a condition with a particular value on a particular variable or when minimizing the variability in a condition by matching the standard deviation of the condition to zero. The equations used to express the costs associated with these single-sample distance constraints are identical to those for the two-sample constraints, only the values of the data points in what have been labeled $\mathbf{x}_{c1}$ are replaced with the user-specified value. For instance, to minimize the group-level difference between a condition, $\mathbf{x}_{c2}$ and a target value, v, simply substitute v for $f(\mathbf{x}_{c1})$ in Equation A.2:

$$O_{valueMIN}(v, \mathbf{x}_{c2}, b, n) = b(|f(\mathbf{x}_{c2}) - v|)^n. \tag{A.5}$$

The same change can be applied to all of the other two-sample distance constraints discussed previously.

Soft entropy constraints. By themselves, soft distance constraints are often sufficient to stipulate the types of relationships among stimuli that are imposed in standard multi-level or factorial designs. However, Baayen et al. (2008) recently criticized these types of factorial designs and argued that they are ill-suited for studying the effects of continuous variables (e.g., studying word frequency effects using low- and high-frequency conditions).

There are two reasons why these designs are problematic. First, they often employ distinct collections of items that are as widely separated as possible on the variable of interest so as to maximize statistical power. Insofar as the distributions across conditions do not overlap, this weakens the degree to which the effects observed at these extreme ends of the continuum can be interpolated to the intermediate values between the conditions that have not actually been

sampled. Indeed, strict adherence to statistical theory dictates that statistical inferences cannot be made to these items, which were not part of the effective populations that were sampled to create the conditions. Moreover, even if this restriction were relaxed, this type of design would require the strong and sometimes questionable assumption that these extremes of the continuum are linked by a particular (usually linear) function which cannot be empirically verified with the items used in this design [e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004, and Brysbaert & New, 2009, report a nonlinear effect of log(frequency) for high-frequency words that would be incorrectly modeled by a linear function]. Second, these types of designs can often be statistically underpowered because they discard the exact data points associated with individual items—represented on a continuum—by collapsing them into qualitative groups in which all items are treated identically and items across groups are treated differently. For instance, two words with frequencies of 1 and 50 might both be grouped together into a low-frequency condition, and two items with frequencies of 49 and 51 might be split into the low- and high-frequency conditions. To the extent that more precise data can actually account for more variance, this kind of falsely dichotomous design results in a loss of statistical power.

After considering these issues, Baayen et al. (2008) concluded that a superior experimental design would involve sampling items that span the whole continuum of values of a given variable and using a form of multiple regression to analyze the results. This procedure has the effect of boosting both the theoretical validity and statistical power of the analyses. The challenge with a regression approach is that variables that are not uniformly distributed will be poorly sampled for some ranges of values and oversampled for others. As a result, statistical analyses can be biased and prevent accurate inferences from being made to at least a subset of the items. Such a scenario is particularly problematic when practical considerations, such as studying patient populations, permit only a relatively small number of items to be tested.

To take full advantage of the statistical framework advanced by Baayen et al. (2008) it is therefore useful to sample uniformly across the range of values for the variables that will be

analyzed. The entropy constraint applies a pressure for stimuli to be distributed in this fashion, either across the range of values for that variable in the entire population or only within the sample. This allows a researcher to differentially accentuate either the external or the internal validity of the experiment, respectively.

The entropy constraint assesses uniformity by examining the degree to which the values of a variable, x, in a given condition are equally distributed in a histogram. The x-axis of the histogram is divided into n bins, where n is by default the number of items and has lower and upper bounds corresponding to the lower and upper bounds of the population or sample, as specified by the user on the basis of the scope of the desired generalization. The y-axis denotes the frequencies of items in the condition that fall within a given bin. An example of such histograms as generated by SOS is presented in the regression-analogue example case in the main text. The entropy, or uniformity, of this distribution can then be formalized as a function that increases in value as the distribution of the variable becomes more uniform. One function that satisfies this criterion is to define a simple measure of entropy, S_{simple}, as a function of the proportion of items, p, in bin i of the histogram, summed across all n bins:

$$S_{simple}(p) = -\sum_{i=1}^{n} p_i \ln p_i.$$

This equation amounts to a simplified version of Boltzmann-Gibbs entropy in which a constant has been set to 1 (Salazar, Plastino, & Toral, 2000). In this case, entropy is maximized when the values of p_i are identical for all i—that is, when values are uniformly distributed in the histogram.[5] This simple entropy equation is then subjected to some minor algebraic adjustments to bring it in line with the other cost functions that have been presented. First, the data on which the other constraints operated were normalized before cost was calculated so that all constraints

[5]In strict terms, there is no attempt to distribute the data uniformly within each bin, so this is true only in the details with large numbers of bins/items. Other, more sophisticated methods could be employed to ensure a perfectly uniform distribution, but their complexity and computational overhead are arguably not justifiable in the vast majority of cases.

would make similar contributions both within and across different optimizations. As it stands, this is not true for the entropy formula that was previously introduced , whose upper and lower bounds change depending on the number of items that are included in the sample. The first step in correcting this problem is to ensure that the entropy function can produce a defined value for every possible distribution of values. However, when the proportion of the total number of items in a bin is zero, the result is undefined. The adopted solution to this problem was to simply add a constant of one when calculating the natural logarithm. Because the natural logarithm produces positive, as opposed to negative, values when applied to numbers greater than one, the resulting equation now produces negative values, but larger (closer to zero) values still reflect a more uniform distribution, as before. Furthermore, the entropy equation can be standardized so that different sample sizes with the same overall distributions of stimuli produce identical entropy values. This is accomplished by dividing the previous equation by the number of items in the condition:

$$S'_{simple}(p) = -\frac{\sum_{i=1}^{n} p_i \ln(p_i + 1)}{n}.$$

Finally, to make the entropy equation more intuitive, it is useful to standardize its range to an interval such as $[-1,0]$. Such a normalization would facilitate the interpretation of intermediate values and make the boundary values for the equation more evident. To accomplish this, it is first necessary to calculate the upper and lower bounds of entropy. The lower bound corresponds to the case where all stimuli fall within a single bin and all of the other bins are empty. In this case, the contributions to entropy for all of the empty bins will equal zero (because $\ln(1) = 0$). Substituting into the existing entropy formula, the lower bound therefore simply corresponds to $\frac{\ln(2)}{n}$. Similarly, the upper bound corresponds to the case where each bin contains 1 of the n stimuli. Substituting into the entropy formula, the upper bound is, therefore, $\frac{1}{n}\ln(\frac{1}{n}+1)$, where the summation across n and the division by n have canceled each other out. Having established the upper and lower bounds, the entropy formula can then be standardized to have a range of $[-1,0]$ via a linear transformation. The standardization moves the upper bound to zero and

divides by the difference between the upper and lower bounds, causing the lower bound value to be shifted to -1:

$$S(p,b,n) = -\frac{-\ln\left(\frac{1}{n}+1\right) + \sum_{i=1}^{n} p_i \ln\left(p_i+1\right)}{\ln\left(2\right) - \ln\left(\frac{1}{n}+1\right)}$$

The cost penalty used to maximize the entropy of a variable is then simply a function of the absolute value of entropy, so that higher entropy values generate lower costs:

$$O_{entMAX}(p,b,n) = b\left(|S(p)|\right)^n \tag{A.6}$$

Soft correlation-matching constraints. Another important consideration when constructing stimuli for multiple/mixed-effects regression analyses is the correlation that may exist between the independent variables. Typically, such a correlation is not desirable because it reduces the degree to which variance accounted for by the overall model can be uniquely attributed to a particular independent variable. This causes an underestimation of the statistical significance of each of the independent variables (by increasing the error variance inflation factor) and, as the collinearity of the independent variables increases, can invalidate the model as a whole (because the coefficient matrix becomes singular; Healy, 1968). Consequently, it is typically desirable to minimize the correlations that exist between the independent variables. In particular, this may be especially useful when re-casting classic factorial designs in terms of regression analyses (Baayen et al., 2008), since this allows for continuous variables — ideally, distributed uniformally across a range, as is possible using an entropy maximization constraint — to be decorrelated without necessitating the creation of full-factorial models where all cells must be included to eliminate such correlations. For instance, such a correlation-matching constraint could be used to stipulate that the length and the frequency values for a given sample must be uncorrelated when creating an analogue of a 2 × 2 factorial design involving those variables, as in the second example case in the main text. Of course, classic factorial designs may further benefit from eliminating such correlations across additional variables that are included as covariates and not as separate factors in the design. More generally, it may also be desirable to match the cor-

relation between two independent variables to a particular value—for instance, to show that an effect reported in a previous study is replicated when such a correlation is present but eliminated when the correlation is removed.

The correlation-matching constraint provides support for these types of scenarios. It allows for users to match the correlation between two variables, $r(\mathbf{x}_{c1}, \mathbf{x}_{c2})$ either within or between two samples, to a prespecified value, v. The cost function used to express this constraint is similar in form to that for the distance minimization constraints, with the addition of two modifying equations that alter the behavior at the boundary conditions when the variance in x_{c1} and/or x_{c2} is zero:

$$O_{matchCorrel}(\mathbf{x}_{c1}, \mathbf{x}_{c2}, b, n, v) = \begin{cases} b(|r(\mathbf{x}_{c1}, \mathbf{x}_{c2}) - v|)^n, & \text{if } \sigma_{\mathbf{x}_{c1}} > 0 \text{ and } \sigma_{\mathbf{x}_{c2}} > 0 \\ 1, & \text{if } \sigma_{\mathbf{x}_{c1}} = 0 \text{ and } \sigma_{\mathbf{x}_{c2}} = 0 \\ 0, & \text{if } \sigma_{\mathbf{x}_{c1}} = 0 \text{ xor } \sigma_{\mathbf{x}_{c2}} = 0. \end{cases} \tag{A.7}$$

Assuming the default values for the weight and exponent, the resulting function is bounded by [0,2] but usually operates between 0 and 1 when attempting to eliminate the correlation between two variables. This is particularly useful because the range of the cost associated with the correlation-matching constraint is therefore quite similar to that of the entropy constraint. Consequently, when both constraints are used simultaneously, which is often likely to be the case, they should each be approximately equally satisfied by default.

Soft meta-constraints. In contrast to soft simple constraints, which operate directly on measurements associated with the stimuli in a condition, soft meta-constraints serve to constrain the differences that exist between constraints themselves. Meta-constraints are useful in some instances when combinations of soft constraints and their respective costs do not correctly express the optimal pattern of relationships that should exist among the conditions. These meta-constraints also serve as a standardized means of re-prioritizing the relative importance of other

constraints that could otherwise require more fine-tuning by the user. Currently, two types of meta-constraints have been implemented: cost-matching meta-constraints and conditional cost-matching meta-constraints.

Cost-matching meta-constraints. The first type of meta-constraint bears a strong similarity to the simple soft constraint used to minimize group-level differences on a particular measure. The main difference is that the measurements contributing to the meta-constraint have been substituted with the costs associated with two other constraints. The primary use of cost-matching meta-constraints is to provide an additional incentive for all constraints to be satisfied to the same degree. This type of behavior usually falls out of the normalization procedure and cost functions naturally but can depend on the distributions of the variables to be matched and the particular set of constraints that must be optimized.

One situation in which a cost-matching meta-constraint may be beneficial is to ensure that a group of soft distance constraints that all maximize differences end up maximizing these differences to the same relative degree. One case in which this does not occur naturally is when the maximal differences that can be achieved for one constraint are possible only by minimizing the differences on another constraint. This is a frequently encountered issue in factorial designs that cross correlated variables, such as when trying to identify items for use in a 2× 2 factorial design that crosses effects of length with effects of frequency. In this case, the correlation between the two variables reduces the extent to which each individual constraint can be satisfied. Instead, maximizing one constraint will tend to also lead to the minimization of the other (e.g., when attempting to select long/high-frequency words, increasing the values on one variable tends to reduce the values on the other). Because costs are typically squared and increasingly large costs result for the same absolute increase in the difference between two conditions, the optimal (lowest cost) set of items may, in fact, be a set that maximizes one difference while actually minimizing the other (despite both constraints stipulating that both differences should be maximized).

This problem can be rectified, however, by stipulating that the two constraints must both be minimized to the same degree. Expressed as a general cost function that minimizes the differences between the costs of constraints O_1 and O_2, this meta-constraint can be written as the difference between the current cost associated with the two constraints with the addition of the standard weight and exponents:

$$O_{MATCH}(O_1, O_2, b, n, d) = 100b\,(|dO_2 - O_1|)^n \tag{A.8}$$

An additional scaling parameter, d (default 1.0), allows for O_1 to be matched to a multiple of O_2 (e.g., so that O_1 should be satisfied only half as well as O_2, in a case where it is more important to ensure that O_2 is well satisfied). Furthermore, the equation also contains a constant multiplier of 100 because, without such a multiplier, the values of the meta-constraint tend to be on the same order of magnitude as the simple constraints over which they operate. Given that meta-constraints are typically applied when the simple constraints themselves are insufficient to adequately characterize the desired optimization, a stronger pressure to satisfy the meta-constraints is required on some occasions. A multiplier of 100 is sufficiently large to robustly provide such a pressure, without overly dominating overall cost as the optimization progresses. Notwithstanding these details, this new cost function bears a strong similarity to the cost function for minimizing differences between two measures in Equation A.2.

With the addition of this constraint, having a very small cost for one constraint and a very large cost for another will be penalized severely, whereas satisfying both constraints to the same extent results in no penalty. This shifts the lowest achievable cost value back toward the state where both constraints are satisfied to an equal degree. The degree of importance placed on ensuring that both of these constraints are equally satisfied can be emphasized by increasing the weight and/or exponent of the equation, as usual. However, the default parameter values typically produce quite a strong pressure to avoid egregious differences between two costs.

A second case that necessitates the use of a cost-matching meta-constraint involves maximizing differences between more than two conditions that span a range of values in a multi-

level design—for example, when creating low-, medium-, and high-frequency conditions that are equally spaced across a range of word frequencies. Without a meta-constraint, the lowest possible cost is achieved when two of the conditions have the same mean frequency and the third is maximally different.

To illustrate this concretely, imagine that word frequencies are bounded between 1 and 11 and that the low- and high-frequency conditions have means equal to these lower and upper bounds, respectively. Without a meta-constraint, the putative medium-frequency condition would actually lower cost the most by also having a mean of 1 or 11, rather than a mean of 6 (we assume a mean of 1 in what follows). This is because the decrease in the cost associated with being very far from one condition [Medium/High cost: $-(10^2) = -100$] combined with with cost of being very close to another condition [Low/Medium cost: $-(0^2) = 0$] results in a lower overall cost (-100) than if the medium condition's mean were located at the midpoint (6) between the low-and high-distributions [$-2(5^2) = -50$]. A cost-matching constraint can overcome this problem by requiring that the cost associated with the difference between the low- and medium-frequency conditions be equal to that between the medium- and high-frequency conditions. The meta-constraint will be minimized when the three conditions are equally spaced $\{2[(-25)-(-25)]^2 = 0$; yielding a total cost of $-50\}$ and maximized with two conditions of equal means and a third condition separated from the first two $\{[(-100-0)^2 = 10000]$; yielding a total cost of 9900$\}$. An example of the successful application of this type of constraint structure is described in the examples section of the main text.

Note that in the previous discussion, the focus has been on instances of matching the degree to which constraints that *maximize* differences are satisfied. This was intentional because, in the case of minimizing differences, the desired behavior is already largely accomplished by the cost-matching constraint: as the difference between the measures being compared decreases, there is a diminishing return on the overall decrease in cost as cost asymptotes toward a lower bound of zero. This behavior can nevertheless be enforced more strongly via a cost-matching constraint,

but it is more likely that these types of constraints will be more useful in the former context. In both cases, however, the two constraints being matched are both either difference minimization or difference maximization constraints; these two types generally should not be mixed. Instead, a conditional match meta-constraint is usually more suitable in those instances.

Conditional cost-matching meta-constraints. Conditional cost-matching meta-constraints are a variant of the basic cost-matching meta-constraints that seek only to match the costs of two constraints when one of the constraints has not been satisfied. The primary use for these constraints is to allow the maximization of differences between variables to occur only if all of the differences to be minimized have been eliminated. This is the case when it is critical that there be virtually no differences on some measurement but there is still a desire to maximize the differences on another constraint. As for the standard cost-matching meta-constraint, this type of constraint satisfaction often occurs naturally but may require an additional pressure in some circumstances—for example, when relatively small increases on the differences to be minimized can lead to very large differences on the differences to be maximized. A pressure that prevents such behavior could be induced by adjusting the weights or exponents of the different cost functions, but these adjustments are likely to be problem specific. Conditional cost-matching meta-constraints are a more general solution to this problem.

A concrete example of a situation that might benefit from a conditional cost-matching constraint would be in the selection of stimuli for an experiment on the effects of word imageability (i.e., how easy it is to picture the meaning of the word in the "mind's eye"; Paivio, Yuille, & Madigan, 1968), using a low- and a high-imageability condition. Generally, effects of imageability are fairly weak relative to other effects, such as word frequency, so it would be desirable to identify items that are maximally different in terms of imageability, provided that there is no difference in frequency. A cost function that allows only differences for a constraint to be maximized, O_{MAX}, conditional on the differences for another constraint being successfully minimized, O_{MIN}, can be written as a function of the costs associated with these constraints and of

the standard weight, exponent, and scaling parameters:

$$O_{condMATCH}\left(O_{MIN}, O_{MAX}, b, n, d\right) = \begin{cases} 100b\left(\sqrt{|(dO_{MAX} - O_{MIN})|\, O_{MIN}}\right)^{n}, & \text{if } O_{MAX} < 0 \\ 0, & \text{if } O_{MAX} >= 0 \end{cases} \tag{A.9}$$

The square-root component of the equation serves to keep the conditional matching function on roughly the same scale as the standard cost-matching function (see the cost-matching meta-constraint description for details on the other parameters). As long as differences on the measurement to be minimized exist, O_{MIN} will be greater than zero, and there will be a pressure to reduce the differences that a simple constraint attempts to maximize. As O_{MIN} is reduced, this pressure is gradually removed, and O_{MAX} becomes a free parameter that the simple constraint will lower as much as is possible. O_{MIN} thus approximates a conditional gating function. The upper bound of 0 when O_{MAX} is greater than 0 is included to prevent undesirable behavior that is possible under this condition. Specifically, when O_{MIN} is greater than 0, without this bound, the desire to minimize the $|dO_{MAX} - O_{MIN}|$ component of the equation would effectively pressure O_{MIN} to maximize the distance between the two conditions in the direction opposite to that intended, until O_{MAX} was as large as O_{MIN}.

Hard constraints

Hard constraints are a special type of constraint that evaluate, in a true/false fashion, whether an item's value on a particular dimension satisfies the constraint. These constraints take precedence over the soft constraints and serve to exclude items that do not meet their requirements from the samples used to fill a condition. Hard constraints can be used if these constraints are known and are of a fixed value for the different variants of an optimization that a user might intend to run.

In the case that a user forces items to be initially placed in a sample that violate these constraints (e.g., by reading in a list of existing items for one condition), hard constraints will cal-

culate a cost penalty such that every time the constraint is violated, a penalty of 1 is added. The total cost associated with hard constraints is accumulated in a special "total hard cost" pool, and the algorithm always attempts to reduce "hard costs" prior to reducing "soft costs." This allows for hard constraints to be optimized first.

Hard bound constraints. The hard bound constraint serves to constrain the value of a particular variable in a condition to fall within an upper or lower bound. This type of constraint can be useful in making minor adjustments to the items that can be included in a condition after examining the performance of an initial optimization without needing to filter out data outside of the software. Hard bound constraints can also be used to speed up the optimization when drawing items for multiple conditions from the same population and trying to cover different ranges of values of a variable. For instance, a low- and high-frequency condition might both sample from the same pool of words. In this case, the exact lower and upper bounds for these conditions may not be known in advance and should, therefore, be left for the optimizer to determine on the basis of how frequency differences can be maximized without violating other constraints. Rather than splitting a population into separate high- and low-frequency subpopulations, a hard constraint could be imposed to help reach this goal by setting some conservative bounds on the two conditions. This would limit the attempted swaps to not include items that would clearly never be part of the "optimal" sample without interfering with the discovery of the optimal separation point between the two groups.

Table A.1: Descriptive statistics for the samples used in Experiment 2 of Morrison and Ellis (1995) and for the SOS optimized samples

Sample		Frequency		AoA		Length		Imageability	
Low Frequency Condition	ME95	6	(3)	3.58	(0.02)	5.04	(0.81)	4.62	(0.72)
	SOS	5	(3)	3.89	(0.91)	5.71	(1.73)	4.80	(0.85)
	SOS[length]	6	(2)	3.86	(0.83)	6.38	(1.56)	4.84	(0.78)
High Frequency Condition	ME95	224	(129)	3.57	(0.83)	5.04	(0.81)	4.62	(0.92)
	SOS	153	(45)	3.86	(0.88)	5.75	(1.85)	4.81	(0.90)
	SOS[length]	161	(47)	3.85	(0.76)	6.38	(1.56)	4.85	(0.82)
(High - Low) Frequency Condition	ME95	218	(130)	-0.01	(0.01)	0.00	(0.00)	-0.47	(0.86)
	SOS	148	(45)	-0.03	(0.24)	0.04	(0.62)	0.02	(0.20)
	SOS[length]	157	(47)	-0.01	(0.18)	0.00	(0.00)	0.01	(0.21)

Note. Standard deviations are in parentheses. ME95 = maually selected stimuli. SOS = stimuli selected by SOS as described in the text and in Subappendix 2. SOS[length] = stimuli selected by SOS with an additional weighting parameter to emphasize matching the conditions on word length. The (High - Low) Frequency section contains the mean pairwise difference and standard deviation of these difference scores across the two samples.

A.8 Subappendix 2: Tutorial and Details for the ME95 SOS Optimization

Here, the process of running an optimization in SOS is described in detail. All optimizations in SOS can be accomplished simply by pointing-and-clicking within the graphical user interface (GUI), which is launched by default in the standalone versions of the software or by the `sos_gui()` command in MATLAB (see Figure A.3). However, users may execute the same commands by writing a simple script file in a text editor, and the example focuses on these text-based interactions with SOS. Creating these scripts is facilitated by having the SOS GUI display the script commands associated with GUI-triggered events. Scripts can subsequently be loaded and run from the GUI or, for users wanting to take advantage of SOS's advanced functionality, within MATLAB.

In this detailed example, SOS is used to create a set of stimuli superior to those discovered in ME95, Experiment 2. This experiment studied the effects of word frequency in the absence of several confounding variables and involved 24 pairs of high- and low-frequency words that differed on frequency while being matched on AoA, length (in letters), and imageability. For use with SOS, the populations from which ME95 selected their stimuli were recreated. Words with Kučera and Francis (1967) frequencies greater than 110 per million made up the high-frequency population, and words with frequencies less than 10 per million made up the low-frequency population.

First, the commands that set up and run a greedy optimization in SOS are reviewed. Then instructions are given for how to calibrate and run a stochastic optimization. In this particular case—and in many cases researchers may encounter—the greedy version of the algorithm performs very well and can, in fact, solve the target optimization problem to a greater extent than can the stimuli reported in ME95. Only in more complex cases or when an extremely optimized set is desired may a stochastic optimization be required. However, to demonstrate the functional-

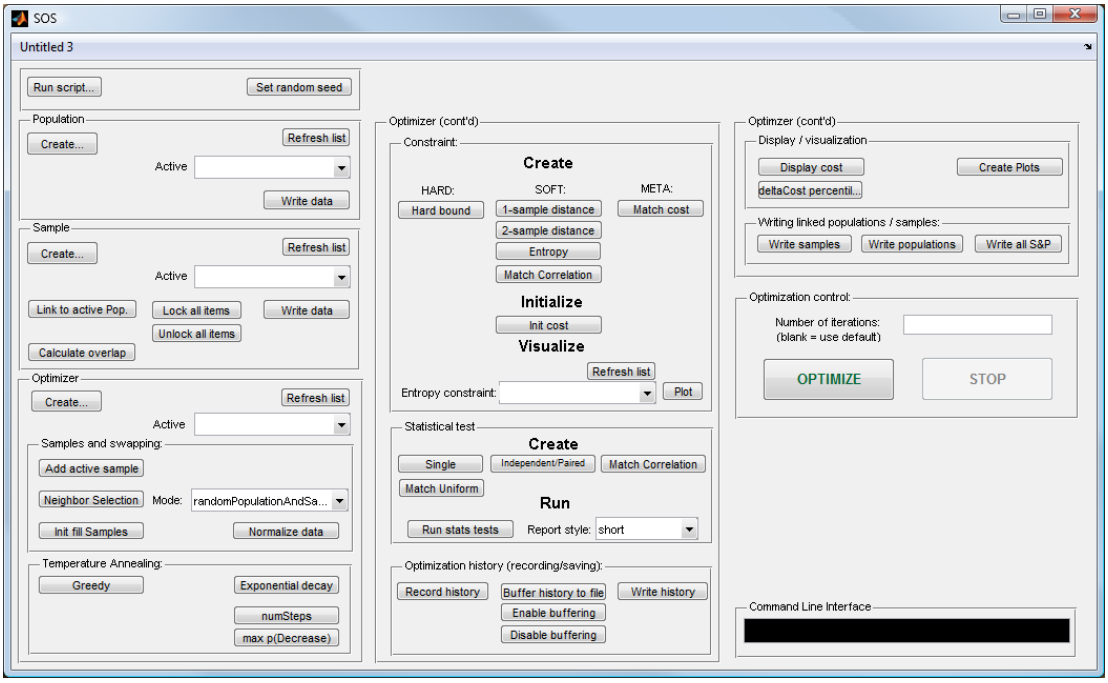


Figure A.3: The main window of the SOS graphical user interface. The process of running an optimization generally flows from top to bottom, left to right.

ity of SOS and the way in which to set up a stochastic optimization, this stage of the process has been included as part of this example. Furthermore, to demonstrate how to calibrate a stochastic optimization from information garnered from an initial run of a greedy optimization, strict limits were imposed on the greedy optimization, so that it artificially ends prematurely.

The script file for this example and all of the realistic examples reported in the main text are available through the online user manual and may be useful as templates for similar optimization problems.

Creating the samples and population. Before starting to work in the SOS software proper, the files that contain the population data need to be formatted in a way that SOS recognizes. In this case, there are two such files: a high-frequency population and a low-frequency population. The items need to be in tab-delimited columns, with one row corresponding to one item. Data sets containing missing values are currently not supported. Each column should also include a

label, or header, at the top of each column so that it can be referred to easily later (although SOS will automatically label the columns if headers are not specified). Additionally, although SOS attempts to auto-detect the type of information present in each column (i.e., words or numbers), users can include this formatting information explicitly by placing a vertical bar ("|") after the column name and either an "s" indicating word/nonnumeric information (strings) or an "f" indicating that the column contains (floating-point) numbers. For example, the first three columns and two rows of the high-frequency population file appear as follows:

```
word|s  AoA|f  KFfrequency|f
act     3.19   237
```

Once the data files have been formatted, the SOS software can be launched, and the data can be read in and used to create populations and samples. To do this, it is necessary to first create the populations from which the optimized stimuli will be drawn. In this case, the two populations are stored in two text files that contain high- and low-frequency words according to the constraints imposed by ME95 and that contain header and formatting information.

```
highFreqPopulation = population('Exp2HighFreqPopulation.txt', 'name', ...
               'highFreqPopulation', 'isHeader', true, 'isFormatting', true);

lowFreqPopulation = population('Exp2LowFreqPopulation.txt', 'name', ...
               'lowFreqPopulation', 'isHeader', true, 'isFormatting', true);
```

The **population** command creates a new population called "'highFreqPopulation"' from the text file, "'Exp2HighFreqPopulations.txt'." Because the columns in this file are labeled with headers and formatting information, the "'isHeader"' and "'isFormatting"' parameters to "'true'." In a different scenario, a researcher may want all lists of words to draw from the same

population. In fact, in the present example, it was possible, in principle, to use a single population of words and then, by using hard bound constraints, imposed the same frequency cutoffs imposed by ME95 (see Subappendix 1). However, the present example strived to follow as similar to that in ME95 as possible, and so, two separate populations of words were created. Additional examples in the main text and user manual describe successful optimizations with this alternative constraint and population structure.

After these populations have been created, the samples that will be optimized can be created.

```
highFreqSample = sample(24, 'name', 'highFreqSample', ...
                    'outFile', 'highFreqSampleGREEDY.txt');

highFreqSample.setPop(highFreqPopulation);
```

Above, a sample of 24 words called "'highFreqSample'" is created using the **sample** command. When the optimization ends, this sample can be saved to a file called "'highFreqSample.txt'." This sample is then linked with the population from which its words will be drawn ("'highFreqPopulation'"). Similar commands are repeated to prepare a low-frequency sample, "'lowFreqSample'".

Creating the optimization. At this point, the optimization itself can be created.

```
Exp2GreedySOS = sos('reportInterval', 100, 'stopFreezeIt', 100, ...
                    'statInterval', 500, 'blockSize', 100);
```

With the **sos** command, a new instance of an SOS optimization with the name "'Exp2GreedySOS'" is created. There are several advanced options that may also be set with this command (e.g., "'reportInterval'," "'stopFreezeIt'," "'statInterval'"; see the user manual for details); if not set,

however, these options take on their default values. In this case, SOS was configured to provide updates and statistical tests more frequently during the optimization process than the default values. Additionally, to later demonstrate the configuration of a stochastic optimization, the number of iterations after which the algorithm will end if cost remains unchanged ("'stopfreezeIt"') was set to be artificially low and changed the block size over which $\Delta cost$ will be recorded when calculating the distribution of cost changes to the same value. As a result, the greedy optimization will end early but will still generate a $\Delta cost$ distribution corresponding to that expected in a minimum—that is, only positive $\Delta cost$ values—during the last block of iterations.

Once the optimization has been created, the samples to be optimized (i.e., "'highFreqSample"' and "'lowFreqSample"') are linked to the optimization (i.e., "'Exp2GreedySOS"') using the **addSample** command—for instance, to add the high-frequency sample:

```
Exp2GreedySOS.addSample(highFreqSample);
```

Adding the constraints. Next, the optimizer must be provided with knowledge of the desired properties of the final samples; in this case, the samples should to differ on frequency and be matched on AoA, length, and imageability. To give the optimizer this knowledge, each of these properties is translated into a constraint and an associated cost function for operationalizing this constraint. Using the **addConstraint** command, the first constraint can be added to the optimizer: maximizing the differences between the lists on frequency.

```
freqConstraint = Exp2GreedySOS.addConstraint( ...
                    'name', 'freqConstraint', 'constraintType', 'soft', ...
                    'fnc', 'orderedMax', 'stat', 'mean', ...
                    'sample1', lowFreqSample, 'sample2', highFreqSample, ...
                    's1ColName', 'KFfrequency', ...
```

```
                    'S2ColName', 'KFfrequency', ...
                    'paired', true);
```

The type of constraint used here is "'soft'" (see the user manual for other options). The "'fnc'" parameter is set to "'orderedMax'" to instruct the optimization to maximize the differences between the samples on the target statistic, which has been specified as the "'mean'" via the "'stat'" parameter. In the present context, this constraint ensures that "'sample1'" ("'lowFreqSample'") will have lower average frequency than "'sample2'" ("'highFreqSample'"). Next, it is necessary to specify the names of the samples and the names of the variables (columns) within each sample to which this constraint pertains. Because in this example it is desirable to make sure that each *pair* of words differs as greatly as possible on frequency (and because ME95 adopted a pairwise selection procedure), "'paired'" is set to "'true'." Advanced parameters (e.g., "'weight'" and "'exponent'") may also be configured using the **addConstraint** command (see the user manual for details). If left unspecified, as in the present example, they will be set to their default values.

Next, the same command is used to create the second constraint: matching the highFreqSample and lowFreqSample on AoA. This involves using the same command to add a new constraint but setting the "'fnc'" parameter to "'min'," as well as changing the variables over which this constraint should operate.

```
freqConstraint = Exp2GreedySOS.addConstraint( ...
                    'name', 'AoAConstraint', 'constraintType', 'soft', ...
                    'fnc', 'min', 'stat', 'mean', ...
                    'sample1', lowFreqSample, 'sample2', highFreqSample, ...
                    's1ColName', 'AoA', ...
                    'S2ColName', 'AoA', ...
                    'paired', true);
```

The remaining two constraints, minimizing the pairwise differences between the two lists on length and imageability, follow this same format.

Finally, three meta-constraints were added to constrain the costs associated with the simple constraints. In particular, these simple constraints consist of one maximization constraint (frequency) and three minimization constraints (AoA, length, and imageability). While the cost values associated with minimization constraints have a lower bound of zero, a maximization constraint's cost does not have a lower bound and can become negative. By virtue of this difference between the two kinds of constraints, one or more maximization constraints can sometimes dominate the cost function and cause other constraints to be more poorly satisfied (as assessed in a pilot optimization without these meta-constraints). This type of problem can be avoided with the use of meta constraints. One type of meta-constraint, "'matchCostNotMin'," effectively prevents a maximization constraint from being optimized until a minimization constraint has been satisfied. In the present example, three "'matchCostNotMin"' meta-constraints were used, one to pair each minimization constraint with the "'orderedMax"' frequency constraint.

```
metaAoAAndFreqConstraint = Exp2Greedy SOS.addConstraint(...
                    'name', 'metaAoAAndFreqConstraint', ...
                    'constraintType', 'meta', ...
                    'fnc', 'matchCostNotMin', ...
                    'constraint1', AoAConstraint, ...
                    'constraint2', frequencyConstraint);
```

To create a "'matchCostNotMin"' meta-constraint, the "'constraintType"' is changed to "'meta"' and "'fnc"' to "'matchCostNotMin'." Then, the names of the two constraints to be matched are. Importantly, when using "'matchCostNotMin"' constraints, the name of the minimization con-

straint must occur as ‘constraint1’. The format for creating the remaining two meta-constraints that equate the frequency constraint to the length and imageability constraints is the same as above.

Now that the optimization and the constraints have been created, it is time to prepare to run the optimization by initially filling the two sample lists. With the **initFillSamples** command, items are selected randomly from the populations to fill the high- and low-frequency samples.

```
Exp2GreedySOS.initFillSamples();
```

Statistical tests. Optionally, a researcher may define statistical criteria for the optimization. If these criteria are met, the optimization will end. Because statistically significant differences or, alternately, non-significant “matches” between samples are often all a researcher wants to achieve (and are the kind of evidence presented to demonstrate that the stimuli were well-suited for studying the target question when results are published), furthermore optimization may be superfluous once reasonable statistical criteria have been met. Moreover, as is discussed in more detail in the user manual, additional processing to the point of reaching a cost minimum may constrain the number of sets of items that can satisfy the constraints, thereby interfering with the generalizability of the results.

In the present example, each constraint is associated with a single statistical test. In the case of word frequency, the aim of the optimization is to ensure that the means of the two samples are statistically significantly different, as follows.

```
Exp2GreedySOS.addttest( 'name', 'freqTest', ...
                    'type', 'paired',...
                    'sample1', highFreqSample, 'sample2', lowFreqSample, ...
                    's1ColName', 'KFfrequency', 's2ColName', 'KFfrequency', ...
```

```
'desiredpvalCondition', '<=',...
'desiredpval', 0.05);
```

Using the **addttest** command, it is possible to associate this statistical test with the current SOS optimization. Because the algorithm will be matching or maximizing the differences between the two lists on a pairwise basis, the type of t-test we will use is a paired samples t-test between the "'KFfrequency"' columns in "'highFreqSample"' and "'lowFreqSample'." These two samples should be be significantly different on word frequency, so the p-value returned by the t-test should be less than or equal to .5 to stop the optimization. The parameters "'desiredpvalCondition"' and "'desiredpval"' together specify this statistical criterion.

In the case of the AoA constraint (as well as the length and imageability constraints), a different statistical criterion is required to infer that the samples are highly similar to one another, and, unlike statistical significance for rejecting the null hypothesis, there is no gold standard p-value for means that do *not* differ. For this example, it is assumed that the two lists are adequately matched if the p-value returned by a paired samples t-test is greater than or equal to .5. This value of .5 is somewhat arbitrary, and other researchers may wish to match to a greater (e.g., p-value of .9) or lesser (e.g., p-value of .3) degree, depending upon the problem at hand and existing standards in their fields. By comparison, in Experiment 2, ME95 were able to achieve "matches" of $p = 1.0$ for word length, $p = .20$ for AoA, and $p = .01$ (i.e., a failed match) for imageability, using manual stimulus selection.

Configuring and running the optimization. The last portion of this example script explains how to run both greedy and stochastic optimizations; the steps for both up to this point are the same. As a first step for all researchers, a greedy optimization is recommended. Many problems are solved with this type of optimization, and it generally takes less time than a stochastic one. Additionally, stochastic optimizations are calibrated with information calculated during a run of the optimization in greedy mode, so a user does not lose any time from beginning a problem in

this way. In the case of the ME95 example, it is shown that, although the greedy optimization finds sufficiently optimal high- and low-frequency sets of words, the stochastic version of SOS performs even better.

To visualize the progress of the optimization, an an optional command is used to create the graphs associated with various critical measures during the optimization, such as the cost and the *p*-values associated with each statistical test.

```
Exp2GreedySOS.createPlots();
```

Greedy optimization.

The optimizer defaults to greedy optimization; that is, it always swaps to items that result in lower overall cost, and the optimization never behaves stochastically. The Greedy optimization can then be run by clicking the 'Optimize' button in the GUI (or typing **Exp2GreedySOS.optimize()** at the command prompt, although the command prompt interface does not currently allow for ad hoc stopping and starting of the optimization by the user). At the beginning of an optimization, SOS calculates the initial value of cost — the cost associated with the items that have been randomly selected to fill the high- and low-frequency samples — and displays the results:

```
Initializing Constraints

    Hard Constraint Total:  0
    Soft Constraint #:1 Cost:  -1.6583 freqConstraint
    Soft Constraint #:2 Cost:  0.62044 AoAConstraint
    Soft Constraint #:3 Cost:  1.1733 lengthConstraint
    Soft Constraint #:4 Cost:  1.01 imageConstraint
Soft Constraint Total:  1.1455
```

```
    Meta Constraint #:1 Cost:  141.3792 metaAoAAndFreqConstraint
    Meta Constraint #:2 Cost:  332.2222 metaLengthAndFreqConstraint
    Meta Constraint #:3 Cost:  269.5026 metaImagAndFreqConstraint
Meta Constraint Total:  743.104

TOTAL COST (soft + meta):  744.2495
```

This display reveals that the three minimization constraints are associated with roughly equal costs before the optimization begins. Because of the cutoffs imposed on the high- and low-frequency lists outside of the optimization, the frequency maximization constraint is already associated with a negative cost value, indicating that the low-frequency list already has lower pairwise frequency than the high-frequency list. Finally, the values of the meta-constraints are by design approximately two orders of magnitude larger than those of the soft constraints; in practice, we have found that this implements a reasonably strong but not overwhelming pressure for these constraints to be well satisfied.

Next, the SOS software begins to print out the values of cost after a specified number of iterations have passed; in this case, we set the algorithm to report the cost value every 100 iterations. Along with this cost information, the estimated time remaining in the optimization (assuming that the maximum number of iterations—in this case, 10,000—has been reached) and the percent of the optimization completed are also displayed. Periodically, SOS assesses whether the user-specified statistical tests have passed their criteria.

```
Optimizing for 10000 iterations
```

```
 Iteration  Cost       % Complete  Elapsed  Remaining
 1)         744.2495   0.01%       1s       1m 10s
 100)       144.85112  1.00%       1s       2m
 200)       76.94003   2.00%       2s       1m 31s
 300)       52.11457   3.00%       2s       1m 17s
 400)       37.08432   4.00%       3s       1m 9s
 500)       32.87713   5.00%       3s       1m 4s
Running all stat tests:
UserHyp:  PASS; highFreqSampleKFfrequency - lowFreqSampleKFfrequency:
t[paired](23) = 15.5669, p = 1.049e-13 p-des:  0.05
m(1) = 162.7083; m(2) = 4.2083 (se=2.0784)
UserHyp:  FAIL; highFreqSampleAoA - lowFreqSampleAoA:
t[paired](23) = -0.40505, p = 0.68918 p-des:  0.5
m(1) = 3.8988; m(2) = 3.9358 (se=0.018688)
UserHyp:  PASS; highFreqSampleletters - lowFreqSampleletters:
t[paired](23) = 1.4968, p = 0.14803 p-des:  0.5
m(1) = 5.875; m(2) = 5.5833 (se=0.039774)
UserHyp:  FAIL; highFreqSampleimagery - lowFreqSampleimagery:
t[paired](23) = -1.1709, p = 0.25362 p-des:  0.5
m(1) = 4.7254; m(2) = 4.8396 (se=0.019902)
```

Over the first 500 iterations of the optimization, cost has decreased rapidly. The frequency difference between the two samples is statistically significant, and the two samples are matched pairwise on length to a greater degree, in fact, than was required. However, the statistical tests reveal that the high- and low-frequency samples are still relatively different on AoA and imageability. Here, an example of a “full” statistical report is shown, but a “short” report is actually the default. The choice of report style may be selected from the GUI or with an optional argument

when creating the SOS object (see the online manual for details).

This greedy optimization continues for only 81 more iterations. At this point, the optimization ends because cost has "frozen" and remained unchanged for our prespecified number of iterations (100). This number was selected to artificially stop the greedy version to allow for the demonstration of the method used to configure stochastic optimization; increasing the number of iterations during which cost remains unchanged before stopping provides an increasingly confident measure that cost has descended into a (potentially local) minimum. In addition to stopping due to frozen cost, other possible termination messages include 'all statistical test passed defined criteria' or 'the [user-specified] maximum number of iterations has been reached [default of 10000].'

When this optimization ends, the high-frequency sample has a significantly higher frequency than does the low frequency sample, paired $t(23)$ = 15.57, $p = 1.01 \times 10^{-13}$. On the other hand, the two lists do not differ significantly on AoA, paired $t(23)$ = -0.41, $p = .69$, word length, paired $t(23)$ = 1.50, $p = .15$, or imageability, paired $t(23)$ = -1.17, $p = .25$. Using even the intentionally limited greedy version of SOS with a small freezing interval resulted in arguably better matches than those selected by ME95 by hand and, it is assumed, in far less time.

Additional insight into the progress of the optimization can be obtained by examining the plots produced during the optimization. The plots associated with the current optimization are shown in Figure A.4. The top plot displays the value of temperature throughout the optimization. Because greedy mode is equivalent to setting temperature to zero for the entire optimization, the temperature plot shows no fluctuations. The next plot displays the value of cost and shows a sharp decrease at the beginning of the optimization followed by very gradual progress towards the final low cost value. The third plot displays the $\Delta cost$ value associated with a swap at each iteration for which data are being plotted and gives a sense of the variability and sign of the cost differences encountered after an attempted swap. For example, large $\Delta cost$ variability indicates high variability in the optimality of the current set relative to the swap set and a large number

of positive $\Delta cost$ values could indicate that the optimization is becoming stuck in a minimum. The fourth plot shows the probability, p(swap), of moving to a neighboring (swap) item, averaged across all of the iterations in a block (by default, of size 1,000). Low p-values on this plot indicate that the algorithm is not moving to neighboring states and provides additional evidence that it could be becoming trapped in a minimum. High p(swap) values indicate that the algorithm is swapping frequently between states, a scenario encountered during the initial stages of the stochastic version of the algorithm or when most items in the population could be selected to satisfy the constraint to the same degree. The last plot shows the p-values associated with the four statistical criteria defined for this example on the basis of the order in which they were created. The 'Names' button shown on the left of the figure displays the full name of the statistical tests in an overlay, not shown. In the present example, the p-value for Test #1 (frequency) remains low throughout the optimization. The other three tests, associated with the minimization constraints, gradually move to the final state where all p-values are greater than .05. The graph shows that Test #2 (AoA) remained high throughout the optimization, while more time was required to minimize the constraints associated with Tests #3 (length) and #4 (imageability).

Stochastic optimization. The next example demonstrates the use of the stochastic version of SOS. To do so, we have used the same example case from ME95 and the same basic script and initial sample of stimuli. In fact, to ensure that greedy and stochastic runs are initialized identically, the **setSeed** function was used to initialize each with the same random number (not shown in the first example), meaning that the high- and low-frequency samples will be filled with the same initial items. Within the script, however, the annealing schedule was changed to use exponentially decaying temperature annealing instead of the default greedy mode (see the user manual for details). Before reviewing how this is accomplished, however, it is necessary to first collect some information to allow for the calculation of *pDecrease*, a parameter in the exponential decay function that specifies the next temperature level as a smaller proportion of the current temperature. The goal here is to identify a value of pDecrease that gradually moves

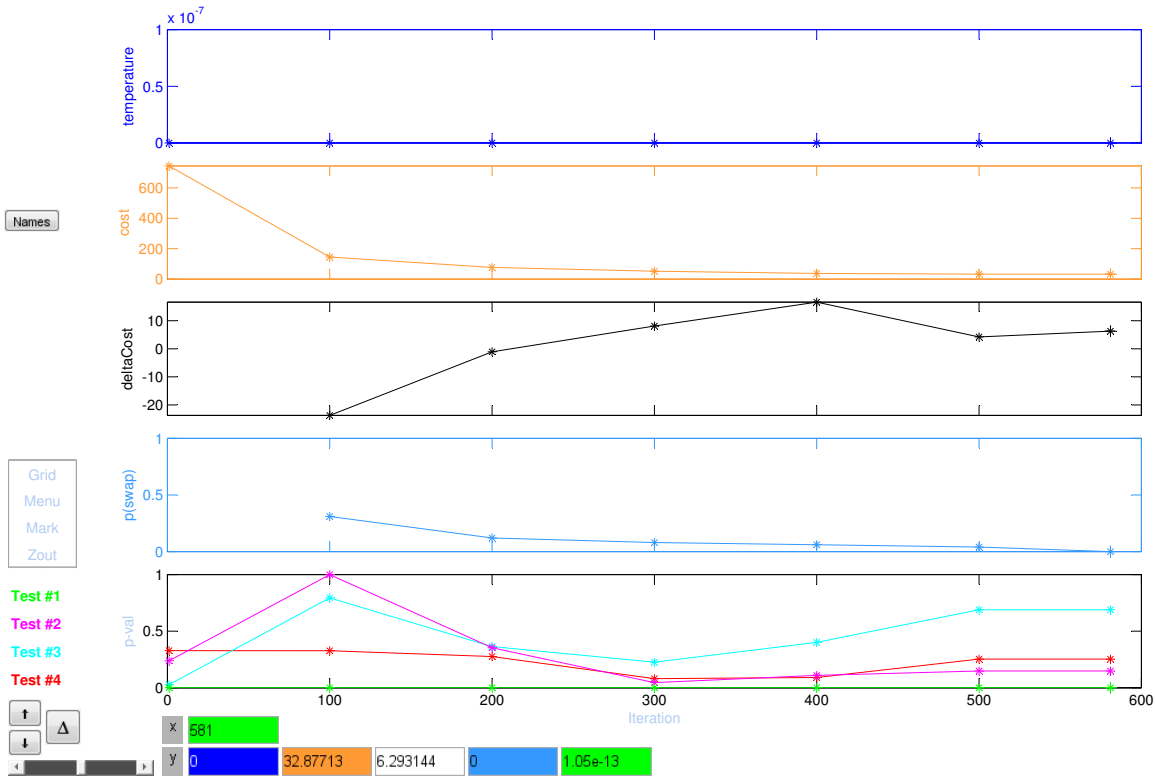


Figure A.4: Plot generated by the SOS software of values for several metrics throughout the course of the greedy optimization. The topmost plot displays temperature, the second plot displays cost, the third plot displays $\Delta cost$, the fourth plot displays the likelihood of swapping to a neighboring item, p(swap), and the last plot displays the p-values for the statistical tests.

from a high temperature, at which swaps occur at random, to a lower temperature that would allow for a small number of swaps to still occur when the algorithm enters a cost minimum. This will allow the algorithm to gradually decrease cost and avoid becoming stuck in a local minimum as it does so.

To determine the appropriate value of pDecrease for an optimization, we offer the following strategy. All optimizations should begin by attempting success through a greedy run of SOS. If a satisfactory solution is not found, users should then extract some additional information from the greedy optimization before beginning a stochastic one. Specifically, in the previous optimization the command **Exp2GreedySOS.deltaCostDeciles** will produce the distribution of $\Delta cost$ values from the last iteration block in the form of a cumulative average (block size is 1,000 by default, but this was changed here to match the artificially low "'stopFreezeIt"' of 100). In the present example, the delta cost deciles are as follows:

```
Deciles for deltaCost in last block:
(100 iterations total; numIt < blockSize in first block)
0:  +0.61125129
10:  +3.87051217
20:  +6.11557827
30:  +6.95036521
40:  +8.88808587
50:  +10.09334703
60:  +11.16903585
70:  +13.83542952
80:  +15.75305910
90:  +19.65537253
100:  +32.39223807
```

```
97.5th - 2.5th percentile deltaCost:  29.10367106
```

Now that the distribution of changes in cost during the final stages of a greedy optimization is known, when it potentially becomes stuck in a minimum in which further decreases are not possible, it is possible to infer a "final" temperature value that will still allow for some cost-increasing swaps to occur for some percentage of iterations when the optimization approaches this minimum. To determine the initial value of temperature, the optimization is run in stochastic mode for one block of iterations; during this time swaps occur at random. After this block completes, the $\Delta cost$ percentiles are displayed once more (output not shown). Armed with this information from greedy and stochastic optimizations, the experimenter is able to configure the exponentially decaying temperature function. Past experience with SOS has shown that setting the upper (initial) temperature of the optimization to the 97.5th - 2.5th percentile of $\Delta cost$ values from the first block of stochastic trials calibrates the algorithm's initial temperature to a reasonable initial value. Specifically, this value is sufficiently high for most swaps to be made effectively at random, but is relatively stable across different optimizations that are initialized with different random seeds (in those instances, the increased variability of using the full range of $\Delta cost$ values leads to substantial variability in the estimates of pDecrease). It is also recommended to initially use the 10th percentile of $\Delta cost$ values from the greedy optimization as the lower (final) temperature. Choosing a smaller or larger percentile for final $\Delta cost$ will lead to faster or slower annealing, respectively, with faster annealing settling more rapidly into a minimum albeit with a greater risk that this minimum is local and not global. Finally, the calculation calls for the number of temperature steps (decreases) that the algorithm will take as it goes from the initial temperature to the final temperature; a minimum value of 10 is recommended (and never less than 3, since fewer steps are effectively no different than starting a greedy search with different initial sets of items). The next code snippet shows the command that calculates pDecrease and its output:

```
expAnneal.maxpDecrease(667.49, 3.87, 10)
```

```
0.402508 is max pDecrease to ensure 10 steps during exp anneal
```

Subsequently, the following command is to indicate that a stochastic optimization is desired and to set pDecrease to the value that was calculated earlier:

```
Exp2StochasticSOS.setAnnealSchedule('schedule', 'exp', 'pDecrease', .402508);
```

This changes the optimization from "'greedy'" to "'exp'" (short for 'exponentially decaying temperature annealing'). With this same command, we have also set an additional parameter that governs the rate at which temperature decreases during the optimization, "'pDecrease'," to the value calculated in the previous section.

Having completed the calibration of the stochastic optimization, the optimization can be initiated. As in the previous greedy example, the optimization begins by calculating the initial value of cost; because the high- and low-frequency samples are filled with the same items in both, the initial cost values will be identical between greedy and stochastic runs. Below are displayed the first 3,000 iterations of the optimization. Note that the maximum number of iterations has been changed here to be larger than the default value to allow the optimization to run for more iterations, as is generally required in stochastic optimizations:

```
Optimizing for 1000000 iterations
 Iteration  Cost        % Complete  Elapsed  Remaining
 1)         744.24945   0.00%       1s       1h 57m 50s
 1000)      3149.89991  0.10%       7s       1h 54m 3s
1000) Annealing Equation calibrated, changing temperature from Inf to 667.4918
```

```
 2000)  1753.88764  0.20%  13s  1h 47m 11s
2000) p(thermEquil):  1.6257e-10 prevBlock m = 1871.7707 (se = 0.79726)
curBlock m = 1679.0243 (se = 0.51208)
 3000)  1368.77390  0.30%  19s  1h 44m 28s
3000) p(thermEquil):  2.8947e-92 prevBlock m = 1679.0243 (se = 0.51208)
curBlock m = 1244.0813 (se = 0.37948)
```

Following the initial block of iterations in which temperature is effectively set to an infinitely high value, the display shows the completion of the calibration of the initial temperature value, and the algorithm will gradually lower temperature throughout the course of the optimization. A simple way of doing so would be to simply lower temperature after a fixed number of iterations have passed. However, it has been found that doing so is often suboptimal because the algorithm will sometimes spend either too much time at high temperatures where swaps are mostly random or too little time at an optimal low temperature that allows for local minima to be avoided reliably. Instead, the algorithm employs a more sophisticated procedure for evaluating when temperature should be lowered, referred to as an "assessment of thermal equilibrium." A full discussion of this algorithm is not necessary at present and is beyond the scope of the present report (see the user manual for details). In essence, however, this procedure lowers temperature only once it appears likely that such a lowering of temperature will not result in the algorithm becoming immediately stuck in a local minimum. This can be approximated by evaluating when cost does not change significantly across subsequent blocks of trials.

Each assessment of thermal equilibrium lists the means and standard errors of the cost values encountered in the previous and the current blocks of trials and the probability that these values are not equal. In the example, thermal equilibrium for this T_0 is not reached until 7,000 iterations:

```
7000) p(thermEquil):  0.75543 prevBlock m = 1431.404 (se = 0.33026)
```

```
curBlock m = 1436.5575 (se = 0.40568)
7000) Thermal Equilibrium Reached - Dropping temperature from
667.4918 to 398.821
```

At this point, temperature is reduced according to the value of pDecrease; in this case, temperature is decreased by approximately 40% whenever thermal equilibrium is reached. Figure A.5 displays the plots generated during the stochastic version of this optimization and reveals additional information about the procedure. During the early iterations, temperature is set to a high value, and so the $\Delta cost$ values vary substantially both positively and negatively because the algorithm is effectively swapping to different items at random. This behavior is directly reflected in the probability of swapping to a neighboring item, p(swap), which is equal to .5 for these early iterations. Similarly, the statistical tests shown at the bottom of the figure vary wildly from iteration to iteration. Eventually, temperature begins to decrease, and, as a result, cost values decrease and become more stable over time. Near the end of the optimization, p(swap) has been reduced to near zero; at this point, the algorithm is effectively swapping only to items that strictly decrease cost. However, even at the end of the optimization, the p-values from the statistical tests appear to be changing; this result suggests that changing even a single item may have a non-trivial impact on the statistics in this particular optimization. After 210,000 iterations, the algorithm stops because all of the statistical criteria have been met. At this point, high- and low-frequency samples have been selected that vary on frequency paired $t(23) = 16.13$, $p = 4.92\text{x}10^{-14}$], but not on AoA, paired $t(23) = -0.68$, $p = .51$, length, paired $t(23) = 0.33$, $p = .75$, or imageability, paired $t(23) = 0.39$, $p = .70$. These results indicate that the stochastic optimization identified a better set of items than did both the greedy optimization and ME95.

Assessing generalizability. In SOS, it is also possible to assess the degree to which the final samples of stimuli are representative of the underlying population of stimuli. If, instead, they are an idiosyncratic subset, it is not possible to make strong claims about the generalizability of any

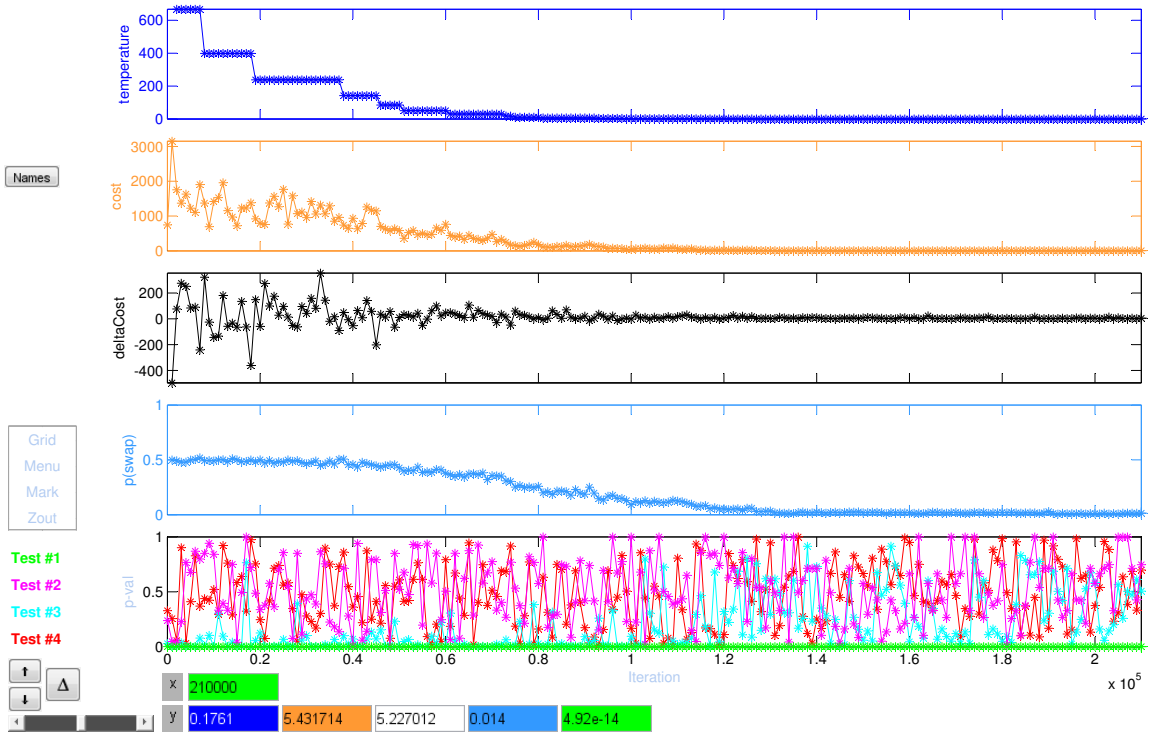


Figure A.5: Plot generated by the SOS software of values for several metrics throughout the course of the stochastic optimization. A more detailed description of the plots is included in the caption for Figure A.4.

experimental effects based on inferential statistics. In the case of the high- and low-frequency word lists, five separate stochastic optimizations were run until each passed the statistical criteria, yielding five lists of high- and five lists of low-frequency items. The "calculate overlap" command (of the form **dataFrame.overlap(sample1run1,sample1run2)** can then be used to determine how similar the resulting lists in each condition were to each other. On average, the high-frequency lists shared 12.92% (SD = 0.09) of their items, while the low-frequency lists shared 7.5% (SD = 0.06) of their items. These numbers suggest that the five optimal solutions were, in fact, quite different from one another and that the algorithm was not arriving at the same, unique solution each time. Thus, any behavioral effects observed with these stimuli are expected to generalize to new sets of stimuli, as well.

A.9 Subappendix 3: Detailed Overview of the Steps in an SOS Optimization

Step 1: Defining the Experimental Conditions

The first step is defining the experimental conditions (different samples of items) in the design. For example, if the purpose of the experiment is to study the effects of word frequency, this could involve creating low- and high-frequency conditions.

Step 2: Specifying constraints and penalties for violating the constraints

Once the experimental conditions have been determined, constraints must be constructed to establish the desired relationships between the different variables, both across and within these conditions. Constraints often consist of a desire to maximize differences on the variables of interest while minimizing differences on confounding variables. For instance, if one were studying the effects of word frequency, the primary constraint might be that two groups of words have different mean frequency values (i.e., low- and high-frequency conditions). A second constraint

might be that the two groups are minimally different in word length.

The issue of assessing how well constraints have been satisfied is more substantial. The basic idea consists of defining a cost or penalty for violations of the constraint. Greater violations lead to increased costs and smaller violations to reduced costs. In the case of manually selecting items, a researcher is assumed to derive these kinds of measures implicitly. In SOS, a cost penalty is operationalized in an objective function referred to as a cost function, or a cost for short.

Many different objective functions could be used to instantiate the idea of cost, as long as they increase monotonically with the degree of constraint violation. For instance, a simple cost function for minimizing differences between two conditions, *c1* and *c2*, on a particular nuisance variable, **x**, is to square the difference between the means of the two conditions on that variable:

$$O_{MIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2}) = (\bar{\mathbf{x}}_{c2} - \bar{\mathbf{x}}_{c1})^2 \tag{A.10}$$

Maximizing the differences between two conditions can be achieved simply by reversing the sign of the cost function used to minimize differences. To enforce a particular rank ordering of the means of the two conditions—for example, that the mean frequency of the high-frequency group is not just different but higher than that of the low-frequency group—an additional term, $s(\bar{\mathbf{x}}_{c1}, \bar{\mathbf{x}}_{c2})$, can be added to the equation corresponding to a simple sign function that equals 1 when *c1* is less than *c2* and -1 otherwise:

$$O_{orderMAX}(x) = -O_{MIN}(\mathbf{x}_{c1}, \mathbf{x}_{c2})\ s(\bar{x}_{c1}, \bar{x}_{c2}) \tag{A.11}$$

This added term ensures that if *c1* is greater than *c2*, larger (positive) costs result. As long as *c1* is less than *c2*, greater differences between the conditions produce smaller (more negative) costs.

By expressing these constraints as costs on the same scale, it becomes possible to assess the overall cost of a particular set of stimuli by calculating the cost associated with each constraint and summing these values. (There is a minor complication in the case of what are refered to as

hard constraints, which are discussed later). To negate the effects of different variables having broader or narrower distributions of values (thus leading to superficially different costs for the same relative differences), SOS calculates cost on the normalized values of each variable rather than on the raw values directly. van Casteren and Davis (2007) noted a similar benefit from normalization in the performance of their algorithm.

In many cases, the use of these two simple cost functions to express constraints related to the maximization or minimization of differences performs remarkably well. An understanding of these constraints at the level of detail presented above may therefore be sufficient for many applications and for acquiring a theoretical understanding of the basic operation of SOS. However, depending on the particular constraints that a researcher wishes to instantiate and the complexity and difficulty of achieving them, it may be useful to modify the default constraints or use other constraints that are more suitable for a given goal. Subappendix 1 provides a detailed description of the full vocabulary of constraints and cost functions currently implemented in SOS, including item-level and group-level constraints, entropy (uniform distribution of items) constraints, correlation-matching constraints, and more. These variations allow for considerable sophistication in creating a wide variety of designs, as illustrated in the example optimizations reported in the main section of the Appendix.

Step 3: Identifying a population and an initial sample of stimuli

Any population of stimuli can be used by SOS as long as the variables to be optimized are expressed on an interval scale. By default, the initial samples for each condition consist of random selections of the desired number of items from the population. Alternatively, some or all of the items in a condition can be pre-specified by the experimenter; further, these items can be 'locked' such that they will not be modified by the optimization. This feature allows for considerable flexibility in the use of the optimizer. For example, an existing set of stimuli can be used as a baseline condition, and new conditions with specified relationships to that baseline can

be created. Alternatively, the optimizer can be used to find replacements for a few experimental items in an otherwise acceptable set of items based on the results of a pilot study. The use of a locked set is summarized in an example case in the main section of the appendix and in additional example cases included as part of the SOS online manual.

Step 4: Search for items in the population and in the sample that could be swapped to better satisfy the constraints

Once the constraints have been specified and an initial sample of items has been drawn, the search for an optimal set of items can begin. In what follows, the simplest form of a search in SOS is presented — a "greedy" optimization of stimuli. Greedy optimizations always seek to swap out existing items in a sample with items that better satisfy the constraints. Although this greedy search has benefits in terms of speed, it is not formally guaranteed to find an optimal or near-optimal set in all situations. Subsequent sections describe how this problem can be overcome by relaxing the deterministic nature of the greedy optimization and allowing swaps to be influenced by random noise—that is, by allowing for the stochastic optimization of stimuli as in a classic stochastic relaxation search (Kiefer & Wolfowitz, 1952).

Representing the optimization problem as movements in state space. To understand how the optimization algorithm operates, it is useful to conceptualize the search for optimal stimuli as a search in *state space* (Rumelhart, Smolensky, et al., 1986). As an example of a simple state space, consider a researcher who has word frequency ratings for a population of four items listed in Table A.2. The researcher's goal is to fill each of the two "item-slots" in the sample with the two items that together have the lowest variability. For simplicity, this constraint can be operationalized as the following simple cost function based on the standard deviation (std) of the sample based on the data in the frequency column, **freq**:

$$O_{MINstd}(\mathbf{freq}) = (std(\mathbf{freq}))^2$$

Additional details on how this constraint can be instantiated in SOS are described in Subappendix 1.

The representation of state space consists of the following. First, each combination of items that can be generated from the population is referred to as a *state*. Each state is represented by a location in space. Each dimension represents the "slot" in the condition that must be filled with a sampled item, and one level on each dimension represents every stimulus that could be used to fill that slot. In the example, the state space corresponds to a 4x4 plane. Two dimensions correspond to the two slots to be filled, and the four levels on each dimension correspond to the four items that could be used to fill these slots. Let it further be stipulated that once an item has been used to fill a slot, it cannot be used to fill another slot (i.e., sampling occurs without replacement). This requirement effectively reduces the plane from 4x4 to 4x3, since the item used to fill the first slot cannot be used to fill the second. An illustration of this procedure is presented in Figure A.6.

Next, an additional dimension is added to represent the cost associated with each state in space. Higher costs are represented as higher elevations, and lower costs as lower elevations. Within this framework, searching for an optimal set of items amounts to moving in this space towards the state with the lowest cost.

In this example, state space is a cuboid, with two dimensions for each of the slots in the sample that must be filled and one dimension for cost. But, for visualization purposes, this

Table A.2: Example 'population' data set containing a single global minimum when attempting to minimize the variability in the frequencies of a 'sample' containing two items

Item	Word	Frequency
1	cow	8
2	pig	10
3	cat	11
4	dog	13

geometric depiction can be altered slightly to allow the first of the "slot" dimensions to represent each of the four possible items directly and the second "slot" dimension to represent the three neighboring items. These neighbors are denoted as n_1, n_2, *and* n_3, in reference to their rank-ordered position in the table once the first item has been removed, as illustrated in Figure A.6. For instance, relative to the items *cow, pig, and cat*, the third neighbor would be *dog*; in contrast, the third neighbor of the *dog* item would be *cat*. Now, the cost for each state can be plotted as a cost surface above the plane formed by the intersection of the different combinations of items, as depicted in Figure A.7. Having created the state space, all that remains is to derive a means of moving to the point with the lowest cost.

	Cow	Pig	Cat	Dog
Cow	-	2.0	4.5	12.5
Pig	2.0	-	0.5	4.5
Cat	4.5	0.5	-	2.0
Dog	12.5	4.5	2.0	-

	N1	N2	N3
Cow	2.0	4.5	12.5
Pig	2.0	0.5	4.5
Cat	4.5	0.5	2.0
Dog	12.5	4.5	2.0

Figure A.6: Upper section: Two-dimensional depiction of the state space formed by each of the different combinations of pairs of items listed in Table A.2. The values in each cell of this table correspond to the cost associated with each combination of items when trying to minimize the variance in a sample of two items. The diagonal is blank due to the restriction that items are sampled without replacement. Lower section: A reduced version of the table from the upper section where the data from the upper triangle has been shifted to the left to merge with the lower triangle. The labels for the vertical axis of this table remain unchanged, whereas the labels for the horizontal dimension now correspond to the neighboring items of the items listed on the vertical axis.

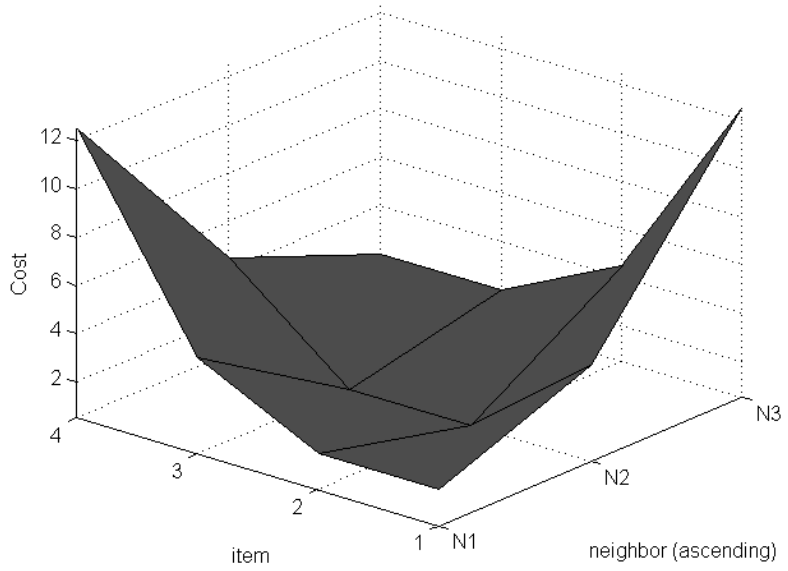


Figure A.7: Three-dimensional representation of the data from the lower section of Figure A.6. The plane at the base of the plot corresponds to the different combinations of items that can be formed from the data listed in Table A.2. The "item" axis corresponds to items 1-4 from the table directly, whereas the "neighbor" axis corresponds to the neighbor items of the item selected on the "item" axis. The "cost" axis represents the cost associated with each state, or combination of items, that could be formed from the population of items being sampled from. Note that the shape of the cost surface is concave when viewed from above and consists of a minimum valley surrounded by hills of monotonically increasing cost. The bottom of this valley corresponds to the global minimum of the state space, and to the pair of items with the lowest variability in frequency values.

Greedy optimization of stimuli. The greedy version of the optimization algorithm is the simplest way to navigate to a state with lower cost. This process begins by calculating the cost of the random location at which the algorithm is initialized. Next, the algorithm randomly selects a condition and an item in that condition for possible replacement. In state space, this corresponds to selecting one of the slots and then allowing the level (item) on that dimension to be changed. The values of all other slots (items in the set) are held constant. A replacement item is then selected from the population (refered to as a *population feeder* in the software) or, additionally, from one of the other samples of items (refered to as a *sample and population feeder* in the software). This candidate replacement item is the level of the selected dimension to which the algorithm would move if the swap actually took place. Having selected a pair of items that could be swapped, the cost associated with the swap is then computed. If the cost of the swap set is lower than the cost of the current set, the swap set is adopted as the current set; otherwise, the current set does not change.[6] This swapping process is then repeated to eventually arrive at a cost minimum where swapping out any item in the sample actually leads to higher cost. Each attempted swap is refered to as an *iteration*.

The greedy version of the algorithm has several useful properties. First, the cost of the swap set can be computed very efficiently by means of a *local update*. In effect, the contribution of the current item in the sample can be removed from affected cost functions, and the contribution of the swap item can be added to affected cost functions; there is no need to effectuate a global update which requires a calculation based on the full set of data. A second benefit of the greedy search is that it can operate without knowledge of the full topography of the state space. It is usually not feasible to collect such knowledge given the impracticality of exhaustively searching

[6] As discussed in detail in Subappendix 1, which focuses on the details of the cost functions themselves, the SOS software implements two types of constraints, "hard constraints" and "soft constraints". The same swapping rule described here applies to both types of constraints, but minimizing hard constraints takes precedece over minimizing soft constraints (i.e., a swap will occur if the cost associated with the hard constraints is reduced even if this also leads to an increase in the cost associated with the soft constraints).

for an optimal set. A greedy optimization, however, can begin its operation from a randomly selected location (i.e., a random sample of items) and lower cost without knowing the full cost topography. The greedy algorithm is somewhat akin to a nearsighted person: it only needs to "see" cost at a neighboring point in state space to decide whether the swap should be made or not. A third advantage of the greedy search is that it rapidly descends in cost space towards a minimum, efficiently focusing its search on the regions of space that are most likely to contain minima. It does so by incrementally ruling out many regions of the cost space for further exploration. Specifically, the algorithm will never consider states of items that require two subsequent swaps that both increase cost relative to the current position in space. For instance, in the four-item example, it would be possible, just by chance, to start the search by selecting the maximally variable sample of *cow* and *dog*. A greedy search would quickly replace these items with a minimally-variable sample containing *pig* and *cat*, and, absent some additional stopping criterion (discussed later), would continue to try (and fail) to replace either *pig* or *cat* with *cow* or *dog*. Relevant to the present discussion is the fact that the algorithm would never re-examine the *cow* and *dog* sample because to do so would require an intermediate uphill step in state space followed by a second attempted step uphill. When dealing with larger populations of items, this property of the greedy algorithm allows more attempted swaps to take place for stimuli that are more likely to lead to the optimal set.

Taken together, the greedy search possesses many of the characteristics of an ideal search algorithm. It is computationally efficient and capable of examining large numbers of sets. Its deterministic cost-reduction rule effectively constrains the number of sets to examine and reduces the time spent exploring less optimal sets than those it has already identified. In many circumstances, the greedy version of the algorithm may, in fact, be sufficient for optimizing stimulus sets. This possibility is always assessed in practice in SOS because the results of a greedy optimization are useful in configuring the parameters of the more sophisticated stochastic algorithm, and so a greedy optimization is usually run first. Nevertheless, the greedy search suffers from a

fundamental limitation, as discussed next.

Greedy optimization's strength is also its weakness—or, the pitfalls of local minima. As noted previously, one advantage of the greedy optimization is that it is only capable of descending to states with lower cost. This feature allows the greedy version to rapidly converge on a set with lower cost and to avoid exploration of parts of the space that are unlikely to contain good solutions. Insofar as the cost topography is perfectly concave, the greedy version is an ideal way of identifying the point with the lowest cost. In practice, however, the cost space is often not perfectly concave. Instead, it may be formed of several unconnected lower-cost valleys separated by higher-cost hills. Since greedy optimizations can only travel downhill, once a valley has been entered, the algorithm will inexorably move towards the valley's floor. The valley floor is a local cost minimum since all of the surrounding states are of higher cost. However, the cost here may still be much higher than the global minimum cost value, present in another valley in the space. Greedy optimization lacks a way to climb over the higher cost barriers that surround the local minimum in order to reach the global minimum.

As a concrete example of this problem, consider the state space that results if the researcher attempting to minimize the variability in a sample's word frequency must select sample items from a slightly different population listed in Table A.3. This population's state space is plotted in Figure A.8. Inspecting this space reveals that the set of items with the lowest variability (*donkey* and *horse*) occurs in one corner of the space comprised of the highest frequency items. This global minimum in cost space is separated from a local minimum formed by the lowest frequency items *mule* and *goat*—which has a relatively low cost value but is still higher than the global minimum—by a barrier of higher cost. This is because there is very high variability in all of the samples comprised of intermediate-frequency items. If a greedy optimization begins at a point within the valley of the local minimum, it will be incapable of reaching the global minimum. A simplified depiction of the local minimum problem is presented in Figure A.9, which roughly corresponds to a multidimensional scaling of the present problem into two dimensions.

Table A.3: Example ‘population’ data set containing a global minimum and a local minimum when attempting to minimize the variability in the frequencies of a ‘sample’ containing two items

Item	Word	Frequency
1	mule	1
2	goat	6
3	donkey	12
4	horse	13

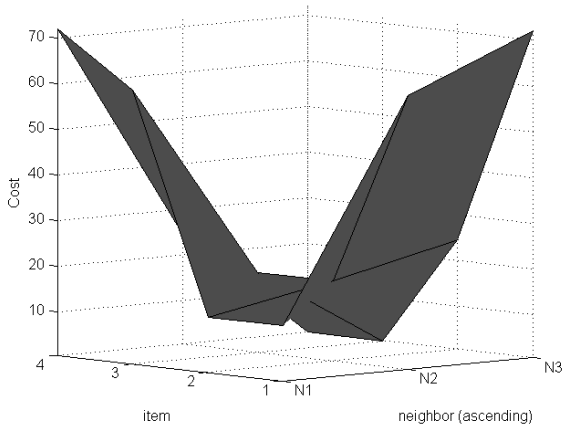


Figure A.8: Depiction of a state space containing both a local minimum (foremost region of the plane) and a global minimum (backmost region of the plane) separated from one another by a ‘hill’ of higher cost. The cost function used to generate the space was identical to that used for Figure A.7, applied to the data in Table A.3.

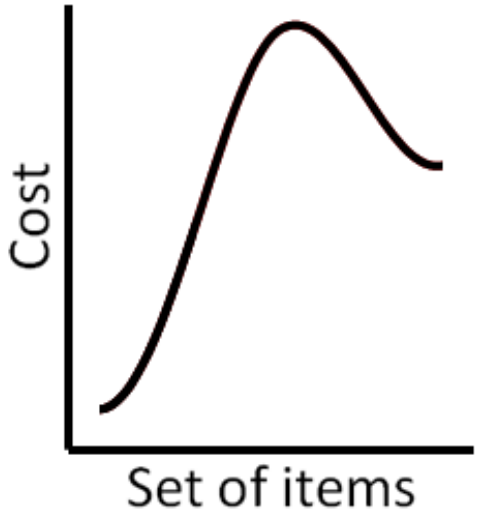


Figure A.9: Approximation of the state space illustrated in Figure A.8 after a multidimensional scaling has been applied to the base plane on that figure to reduce it to a single dimension.

Local minima are likely present in the cost topographies of many different optimization problems, and their presence and locations are rarely known precisely because the entire cost topography usually cannot be computed. As a general guideline, the likelihood of becoming trapped in local minimum increases as sample size increases, as population size decreases, and as the number and complexity of the constraints increases.

Stochastic optimization of stimuli. The problem of local minima precludes greedy optimization as a general method for discovering optimal stimuli. However, the idea of constraining the search to the regions of state space that are most likely to contain a minimum often leads to succesful optimizations that are superior to the sets discovered by a purely random search of the space. By interpolation, employing a form of stochastic optimization search, which involves both a greedy-like bias on swap decisions combined with a random noise term, is a logical solution to these problems (e.g., Kiefer & Wolfowitz, 1952). In detail, the implementation of a stochastic algorithm consists of replacing the deterministic cost-reduction rule from the greedy optimization with a stochastic choice that can be biased to keep or swap items on the basis of the

relative cost difference, $\Delta cost$, between the current set and the swap set (Hinton & Sejnowski, 1986). Formally, this choice is made by passing $\Delta cost$ through a sigmoid function with a temperature parameter, T, which has a lower bound of zero. Temperature specifies the steepness of the slope of the linear portion of the sigmoid in the range surrounding zero and, in conceptual terms, determines the degree to which $\Delta cost$ can bias the probability of staying with the current set or moving to the swap set:

$$p_{swap}(\Delta cost, T) = \frac{1}{1 + e^{\frac{-\Delta cost}{T}}} \tag{A.12}$$

The role of the temperature parameter is to determine how strongly differences in cost between the current state and the swap state are able to bias the likelihood of making a swap. More formally, the likelihood of swapping to a particular state follows a Boltzmann distribution and is determined by the $\Delta cost$ of the swap state relative to the current state, as well as the temperature. Figure A.10 illustrates this relationship in several different sigmoid functions generated with various temperature parameters. When temperature is very high, the sigmoid collapses into a line with a slope of zero. At such high temperatures, even large cost differences are unable to bias the likelihood of making a swap, and swaps are made at random. When temperature equals zero, the sigmoid function collapses into a step function and the algorithm makes identical swap decisions as in the greedy algorithm. At intermediate temperatures, however, the likelihood of making a swap is biased towards lowering cost without necessitating that all steps always lower cost—all possible swaps are associated with a non-zero likelihood and could potentially be made. As a result, as long as temperature is greater than zero, this algorithm is guaranteed, in a formal sense, to be able to move to the global cost minimum, provided that the optimization is run for a sufficiently long period of time.

Although this formal guarantee offers a basic assurance that the algorithm is capable of finding an optimal set, in practice, selecting an appropriate temperature for a given optimization problem poses a challenge. If the temperature is too high, the algorithm will effectively be per-

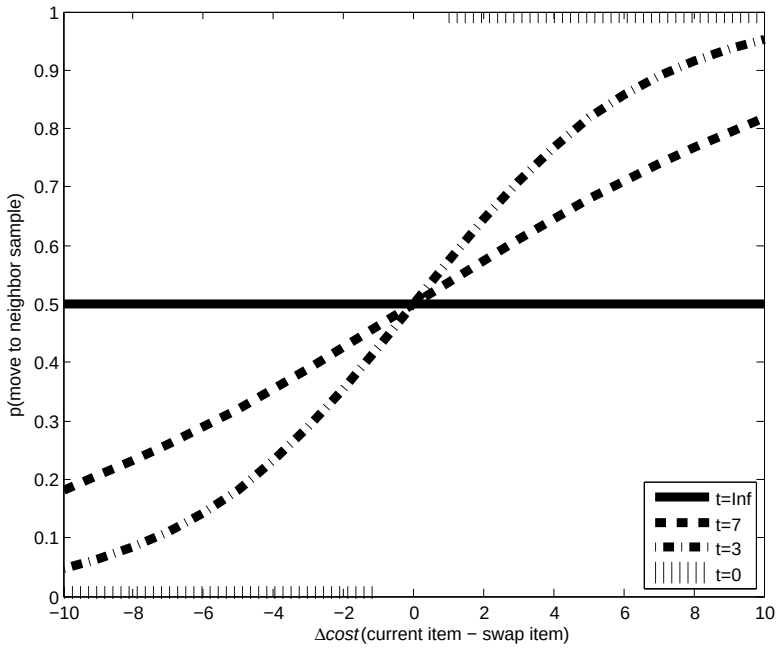


Figure A.10: Depiction of the likelihood of moving to a neighboring sample in state space by swapping two items based on the cost difference, $\Delta cost$, between the two items that could be swapped, for a range of temperature values. Inf = infinite.

forming a random search of state space. If temperature is too low, then the algorithm performs a greedy optimization. As discussed previously, neither of these options are ideal. It is therefore useful to consider the lower and upper bounds for the range of temperatures that allow stochastic optimization to operate effectively.

To determine the effective upper bound on temperature at which nearly all swaps are random, it suffices to allow the algorithm to explore the different states in state space completely randomly (e.g., by effectively setting temperature to ∞) for a block containing some prespecified number of iterations (1000 generally works well). As the algorithm randomly explores the space during this block, the $\Delta cost$ values that result from each attempted swap are recorded. At the end of the block, temperature can then be set to a high enough level that even the most extreme $\Delta cost$ values exhibit only a modest bias on the likelihood of swapping two items. After experimentation with several different values, it has been found that setting the upper bound on temperature to the range of $\Delta cost$ values that encapsulates 95% of swaps (specifically, the 97.5th - the 2.5th percentiles) during this block of random swaps is an effective value. This is a reasonable approximation of the upper bound of temperatures which allow $\Delta cost$ to affect the likelihood of making a swap which is also fairly consistent across optimizations (the range of $\Delta cost$ values being far more variable across different optimizations).

Determining a lower bound on temperature involves running the greedy version of the algorithm until it becomes trapped in a minimum. At this point, it is possible to infer that the reason the algorithm became trapped was because all of the swaps that were attempted while in the minimum were of a higher cost than the current state. Without knowing whether this minimum is local or global, a conservative assumption would be to assume that it is a local one and that the algorithm would have been able to escape it by making one or more uphill swaps. To escape from the minimum, it would therefore be necessary to raise temperature to a point at which a small subset of the attempted swaps are substantially influenced by the stochastic process. This change would allow the algorithm to move uphill to these items and potentially escape the minimum.

By recording the $\Delta cost$ values across the last block of iterations (typically 1000) of a greedy search, the sampling distribution of $\Delta cost$ at various percentiles can be examined (all of which, by definition, are greater than zero, since no swaps were able to further reduce cost). Then, temperature could be set to the $\Delta cost$ for a given percentile to allow all of the attempted swaps below that percentile to have a substantial likelihood of being influenced by the stochastic process and thus able to lead to uphill movements. The specific percentile that is selected as the lower-bound on temperature reflects a trade-off between how greedily the algorithm behaves and how likely it is that the algorithm will be able to avoid a local minimum. After running a wide variety of optimization problems, it has been found that selecting the $\Delta cost$ from the 10th percentile has generally been a good starting value, although a larger value (e.g., from the 50th percentile) may lead to a more optimal set, albeit after many more iterations.

Adding a temperature annealing schedule for fast, reliable optimizations. Having established the lower and upper bounds for temperature, the next question to ask is whether the temperature used in an optimization should be nearer to the upper bound or to the lower bound. In an ideal situation, the goals of the stochastic optimization algorithm should be to allow for 1) stochastic movement that travels uphill in cost space when stuck in a local minimum and 2) rapid, greedy movement once the algorithm has entered the valley that contains the global minimum. A gradual annealing process[7] which lowers temperature over time achieves both of these goals (Kirkpatrick et al., 1983). This allows for an early period of exploration of a variety of different states and gradually eases into the global minimum (Hinton & Sejnowski, 1986).

To understand how the annealing process accomplishes this feat, consider the ball-in-a box metaphor of temperature annealing outlined by Hinton and Sejnowski (1986). Imagine that a

[7]Movellan and McClelland (1993) note that there is a slight distinction between simulated annealing (Kirkpatrick et al., 1983) and simulated sharpening (Akiyama, Yamashita, Kajiura, & Aiso, 1989), and the implementation of annealing in SOS is actually closer to the latter. However, it is expected that the annealing metaphor is more intuitive and have therefore overlooked this distinction in the present work.

model of the cost topography shown in Figure A.9 was placed in a box. Next, a ball is placed into the box. The ball's location corresponds to a particular point in state space. The box is then sealed so that the topography cannot be seen. Now, imagine that someone who has not seen the topography is instructed to move the ball to the lowest point in the box—the global minimum. Without knowing the topography, this person must therefore rely on two known principles and the gradation between them to move the ball to the global minimum. If the box were shaken vigorously, the ball would bounce about and potentially visit every possible location (state) in the box (stochastic search with high temperature). If the box were set still, the ball would roll downhill to the bottom of the closest valley (greedy search).

By starting off by shaking vigorously and then gradually shaking the box less and less, it will be possible to move the ball to the global minimum. Shaking the box vigorously initially allows the ball to avoid becoming trapped in any local minimum. By incrementally decreasing the vigor with which the box is shaken (by decreasing temperature), a useful property emerges: at some point, the shaking will be providing the ball with enough energy to bounce out of a local minimum, which is by definition shallower than the deeper global minimum, but insufficient energy to move uphill from the global minimum back into the local minimum. It is therefore possible to ensure that the ball effectively becomes contained within the global minimum to an arbitrarily high probability.

In cost topographies that contain many hills and valleys, gradually lowering temperature also ensures that the largest hills are excluded from being visited before the smaller hills. This results in the early elimination of the greatest violations of the constraints. The algorithm then proceeds to deal with the more fine-grained and subtle topographic differences that exist amongst the remaining states once these larger differences have been ruled out.

To implement these ideas formally, an exponential decay function is employed. This equation has been shown to be successful in the annealing of connectionist networks, which are close relatives to SOS (Hinton & Sejnowski, 1986) and represents the optimal annealing function when

such networks are allowed to learn their own annealing functions via gradient descent (Ackley et al., 1985; Plaut, Nowlan, & Hinton, 1986). In the decay function, temperature, T, is determined by three parameters: The first parameter is the initial temperature, T_0, which corresponds to the $\Delta cost$ at the (97.5th – 5th) percentile of the sampling distribution of $\Delta cost$ during the initial block of completely random swapping. This temperature exponentially decays towards 0 as a function of the second parameter referred to as the decay rate, λ, and the third parameter, which specifies how many times temperature has been decreased (these are refered to as temperature steps, with the number of steps denoted as *nSteps*). Each step typically contains at least 1000 iterations, for reasons discussed in detail later.

$$T = T_0 \cdot e^{-\lambda \cdot nSteps} \tag{A.13}$$

Plots of the temperature values for a given number of steps were generated using different values of the λ parameter (and a corresponding *pDecrease*, discussed later) and are presented in Figure A.11. Note that higher values of λ lead to more drastic temperature decreases at each step, whereas lower values of λ lead to more gradual temperature decreases. By varying λ, it is therefore possible to trade off making a rapid transition from random to deterministic behavior—which leads to faster settling into a minimum—and making a slow transition—which ensures that the algorithm does not get stuck in a local minimum.

Two factors complicate the intuitiveness of the λ parameter in the previous equation. First, λ is an unbounded parameter which affects temperature in a nonlinear fashion. Second, whether a given λ leads to a very aggressive or very gradual annealing schedule is determined by the specific optimization problem at hand. The latter point is the result of differences in the "elevations" of the minimum and maximum cost values across optimization problems, leading to different effective temperatures at which the optimizer will be making swaps at random versus only making swaps that lower cost. We have attempted to address these issues so that users of the SOS software can more readily select an appropriate aggressiveness for the decay function

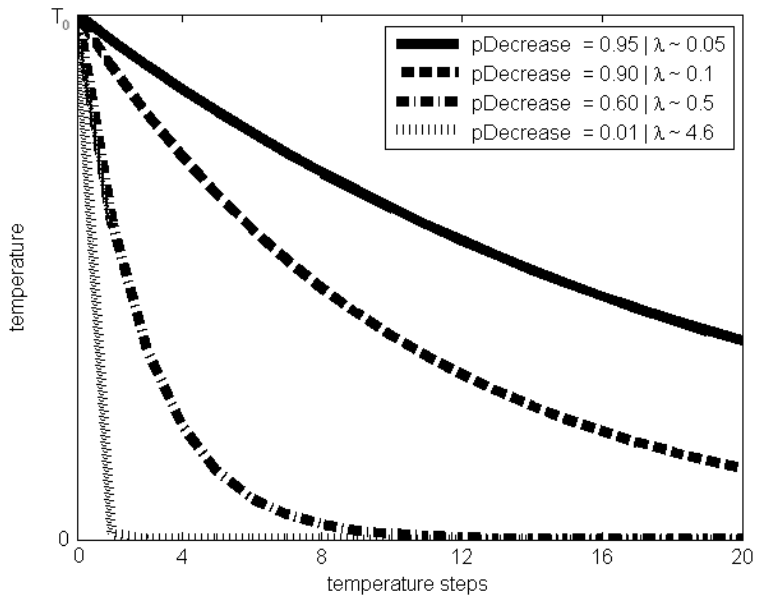


Figure A.11: Depiction of the temperature values after a given number of temperature steps away from an initial temperature, T_0, as a function of a range of pDecrease and λ values.

as follows: Instead of specifying λ directly, the current implementation of SOS allows for the indirect specification of λ via several other parameters. This procedure is explained in detail in the next section, although in practice helper functions are used to conduct these calculations and thus very little user interaction is needed.

The parameters needed to calculate λ indirectly are the initial temperature at which the system shows random behavior, T_0, the final temperature at which only a pre-specified portion of swaps will be made randomly (for the rest, the algorithm will basically operate in greedy mode), T_{final}, and the number of steps, *nSteps* that are needed for temperature to decrease from T_0 to T_{final}. The first two parameters correspond to the upper and lower temperature bounds outlined previously. The *nSteps* parameter then determines how quickly temperature passes from the upper bound to the lower bound. The choice of the number of steps to use reflects a compromise between minimizing run-time and avoiding local minima. At the very least, three steps should be used, as the first step will correspond to a random search and the last step will be similar to

a greedy search. As a guideline, it has been found that at least 10 steps should be used in most optimizations, and more steps tend to lead to better optimizations. Once T_0, T_{final}, and *nSteps* have been determined, these values can be substituted into Equation A.13 to determine λ. Having determined λ, the fully specified exponential decay equation can then be used to calculate the proportion of the temperature that will be subtracted from the temperature at step *n* to determine the temperature at step *n+1*. This proportion is refered to as pDecrease.

In practice, with the help of this indirect equation and the helper functions in SOS, all a user must do to configure the decay equation is as follows: First, run the optimizer in greedy mode to see if it can find a sufficiently optimal set. If it fails, the $\Delta cost$ value from the desired percentile (e.g., the 10th) should be recorded, and the user can pass this information, along with the desired number of steps (e.g., 10), to a helper function that calculates pDecrease. Once pDecrease is set, the decay equation (and by proxy, λ) can be configured automatically, and a stochastic optimization can begin.

Using a proxy of thermal equilibrium to assess when temperature should be lowered. Having decided to use a gradually decreasing temperature value, one question still remains: When should temperature reductions occur? After some experimentation with other schedules for decreasing temperature, it has been found that it is best to do so only once sufficient evidence has accumulated so as to reliably estimate that the position of the current set of stimuli in state space no longer falls within a local minimum from which it could not escape at lower temperatures. This can be estimated by considering when the optimization at a particular temperature has reached a state referred to as *thermal equilibrium*.

Thermal equilibrium is the point at which the probability of swapping to a given item has stabilized at a particular value across all items. To understand how this measure could serve as a basis for assessing when to lower temperature, consider once more the simple state space with a single local minimum presented in Figure A.9. Assume for the moment that temperature has been set a little above the level needed to move from the valley containing the global minimum

to the valley containing the local minimum. Also assume that the optimization has been run for a sufficiently long amount of time to determine the probability of swapping a particular item into the sample set. At this point in time, the probability distribution for the sets in the valley of the global minimum will be higher than those in the valley of the local minimum since $\Delta cost$ should be substantially biasing the likelihood of making swaps into and out of each valley. Nevertheless, in both cases, these probability distributions will contain values that are non-trivially above zero.

Next, imagine that temperature is lowered so that it is just possible to leave the valley containing the local minimum and enter the valley containing the global minimum. Once in the valley containing the global minimum, the relatively low temperature will provide a substantial bias to continue moving downhill in this deeper valley. As cost descends further, the likelihood of making a swap back into the valley containing the local minimum gradually moves towards zero. If the optimization continues at this new temperature for some period of time, the probability distributions for swapping to each item will eventually stabilize once more. The new probability distribution will differ from the old one, however, because some of the probabilities (i.e., those contained in the local minimum) will be lower in the new distribution, whereas others (i.e., those in the global minimum) will be higher. At this point, temperature could be lowered again to hone in on the global minimum more quickly.

Of course, if the temperature is still substantially above the level needed to probabilistically exclude a portion of the state space containing a local minimum (e.g., if temperature were still an order of magnitude above the temperature value described in the previous example), these probability distributions might not differ substantially before and after temperature is changed. Thus, one cannot simply compare the probability distributions before and after the change in temperature to infer if temperature can be lowered. Instead, it is necessary to estimate the probability distribution repeatedly at a particular temperature and lower temperature only when consecutive probability distributions are nonsignificantly different from one another. When this is true, thermal equilibrium has been reached. To maximize the accuracy of this inference and to ensure that

the distributions have, in fact, stabilized, it is necessary to try many attempted swaps before each assessment of whether thermal equilibrium has been reached so as to allow the distributions to become stable.

In principle, examining the stability of the probability distribution of swapping to different items is a good way to assess thermal equilibrium. In practice, however, it can take very large blocks of iterations before the probabilities of swapping to each item stabilize. To reduce run time and memory use, it is far more efficient to monitor a proxy of the probability distribution—the distribution of cost values. This measure provides a summary of the probability of swapping to a given item based on the cost value associated with swapping to that item, averaged across all items, and is already calculated as part of the optimization. The cost distributions for two subsequent blocks of trials can also be readily compared using t-tests, such that thermal equilibrium is said to have been reached once a standard statistical criterion has been exceeded (e.g., $p > .05$).[8]

Step 5. Evaluating how well the constraints have been satisfied and stopping the optimization

The search for an optimal set of stimuli can be terminated automatically when one of several conditions is met. The first consists of passing statistical tests specified by the user. These tests assess when the constraints have been satisfied to a desired threshold. There are currently three main types of statistical tests available for use in SOS: one- and two-sample t-tests (paired and independent), correlation-matching tests, which test whether a correlation in the sample matches a target value (typically 0 to eliminate correlations between independent variables and boost power in a regression), and a Kolmogorov-Smirnov test for assessing the uniformity of a distribution. These parametric statistical tests are approximate analogues to the distance minimization and maximization constraints, correlation-matching constraints, and entropy constraints, respec-

[8]Empirically, the distribution of cost values typically deviates somewhat from some of the assumptions underlying a t-test (e.g., normality). However, the goal of this comparison is to provide a quick and relatively reliable means of assessing when thermal equilibrium has been reached, and t-tests are more than adequate for this task.

tively. For instance, if the goal is to make two conditions maximally different on word frequency, a user could specify that a t-test on the word frequencies across the two conditions should return a t-statistic associated with a p-value that is less than .05. Similarly, if another goal is to match the conditions on word length, the user could specify that a t-test across the word length data in the two conditions should return a t-statistic associated with a p-value greater than .5, which reflects a failure to reject the null hypothesis. Different critical p-values may be desirable both depending on the importance of satisfying different constraints during the optimization and the specific parameters of the constraint in question. For instance, the statistic used to assess whether a sample correlation matches a target (population) has larger confidence intervals when testing whether two near-zero correlations are matched than when two large correlations are matched because, by definition, there is less unexplained variance in the latter case. The test of uniformity of distributions is also capable of detecting small deviations from uniformity that may not be particularly important in many situations. Using a smaller critical p-value (e.g., .001) when testing whether a sample distribution is uniform and inspecting the plots of the distributions (which can be accomplished via the SOS software) when the test is not passed to determine if they are due to a failure to distribute items across particular ranges is therefore recommended.

The optimization terminates once all of the statistical tests exceed the user-specified criteria. Because these probabilities generally do not change drastically after a single iteration, statistical tests are only run periodically (the default is every 1000 iterations). Note, however, that blind reliance on these statistical tests may in some cases be misleading because statistically significant differences are sometimes detected for extremely small effect sizes (e.g., a word frequency difference of 0.1 can be significant between two conditions because the variability within the conditions is .001). This is most frequently the case in the context of pairwise minimization constraints, which in minimizing pairwise differences also minimize the variability in these differences, which in turn leads to an apparent lack of improvement or worsening of the match. Instead, failures to satisfy the statistical tests when the optimization ends for one of the other

reasons discussed next typically warrants inspection to confirm that the statistical differences between the conditions are theoretically important. If appropriate, a user can also opt to set a minimum difference threshold when configuring the tests. This allows the statistical tests to "pass" once the difference between the two conditions has decreased below this bound, independent of statistical significance. A special print-out is displayed when tests pass based on the threshold instead of the statistical criteria.

The second condition that can cause the termination of the search is when the sample has reached a "frozen" state. That is, cost has not changed for a specified number of iterations (the default is 10,000). Note that swaps may still be occurring between items with the same impact on overall cost (i.e., items for which a swap would generate a $\Delta cost$ of exactly zero). Frozen states should generally only occur if the search has become trapped within a minimum and further descents in cost are no longer possible. Using a very gradual annealing schedule and allowing the optimization to reach a frozen state in stochastic mode should therefore correspond to finding a global minimum in the search space. If the only goal is to identify the set(s) of items that best satisfies the constraints, then this is the preferred mode for terminating the search. However, letting the optimization reach a frozen state may not necessarily be desirable if the goal is to use the selected items to make broad generalizations to other items, as discussed in the next section.

The third stopping condition is when a pre-specified number of iterations has elapsed, which prevents the optimization from running for an indeterminate amount of time. Generally, users should set this value to as high a value as they are willing to wait for an optimization to finish. The software defaults to a relatively small number of iterations (10,000) so that users can confirm that their optimization procedure is roughly behaving as it should before increasing the value of this stopping condition. A value of at least 500,000, and possibly into the tens or hundreds of millions is recommended depending on the complexity of problem being run, the speed of the computer, and the patience of the experimenter. If this stopping condition is reached, the experimenter can then consider whether the resulting set satisfies their constraints or if an additional pressure (e.g.,

a more aggressive annealing schedule) is needed to encourage the optimization to reach a frozen state before this amount of time has elapsed.

Step 6. Assessing the degree to which the items in a sample represent the broader population

Although one of the principal goals of many researchers is to infer the results of an experiment using a particular sample of stimuli to a broader population, present practices in stimulus selection place virtually no emphasis on assessing how representative a sample is of the population. Consequently, the final step in the optimization procedure does not have a clear analogue in many experimenters' stimulus selection process. However, constraints imposed on selecting experimental items can effectively restrict the sampled items to a very unusual subpopulation or even a uniqe set of items. At that point, stimuli should be treated as fixed as opposed to random effects and cannot be used for the purpose of statistical inference to other stimuli (Hino & Lupker, 1996). Clearly then, failing to assess the generalizability of stimulus sets is a critical, yet overlooked, aspect of experimental design. Such a state of affairs is quite paradoxical given the strong emphasis that is placed on reporting statistical inferences that generalize across items, as it is unclear whether such inferences are appropriate (e.g., by-item and by-subject analyses, F-max, Clark, 1973; mixed-effect regression, Baayen et al., 2008).

Automated optimization procedures do, however, offer a means of assessing generalizability. The simplest of these methods involves a direct measurement of how many different items could have been used to satisfy the constraints imposed by the user (i.e., the size of the effective population). This is achieved by running the optimization several times and comparing the sets that satisfy the constraints. If many different sets of items can be produced which all satisfy the constraints, there is a basis to generalize the particular set of findings to a broader population of items; otherwise, the sample cannot be used as a basis for statistical inference to other stimuli. As a general guideline, it is recommended that if the average overlap between sets of stimuli is

below 25%, generalizability is likely not an issue. Researchers should typically be concerned about failures to generalize when the average overlap exceeds 50%, although the details of the optimization problem and indended generalization may lead a researcher to adopt somewhat different thresholds.[9]These additional searches may not even take a noticeable amount of time to run on a modern multi-core processor, as many optimizations can be run simultaneously.[10]

The procedure outlined above is likely suitable for assessing generalizability in the majority of cases, where the main objective is to assess whether a relatively broad population of items meets the constraints. However, in some cases a researcher might be particularly concerned with garnering some stronger assurance that the actual population is very well represented in the sample of items. In this case, soft entropy constraints can be used to ensure that the breadth values of a particular variable across the entire population are sampled. This is arguably the best way of ensuring the external validity of the sampled items but may make it difficult to simultaneously satisfy other constraints.

The overall optimization process thus explicitly reflects a trade-off between internal and external validity, and different investigations may warrant different trade-off schemes. At the very least, SOS makes the assessment of this trade-off and of the effects of re-weighting the impor-

[9]It is worth noting that in the details, the average amount of overlap may be too coarse a measure for making a fine-grained inference about generalizability, because it may be possible that a very small set of items which do overlap across the different samples are necessary to satisfy the constraints in all of the optimizations. A simple way of assessing this possibility is to save the residual population after an optimizated sample has been discovered (methods of doing so are described in the user manual) and reruning the optimization sampling from the residual population. Alternatively, and more rigorously, users could essentially set up different variants of the same optimization and have them all run simultaneously while sampling from the same population. The latter is in fact the more formally appropriate method of ensuring the generation of equally comprable samples as all of the samples effectively had equal access to all of the items. However, given current practices in the field, any assessment of generalizability would represent a substantial advance.

[10]This can be accomplished either by launching the SOS executable multiple times or by running SOS using a MATLAB installation that includes the Parallel Computing Toolbox. Note that a different random seed is needed when initializing each optimization.

tance of these two factors a relatively easy and effortless task. This possibility, in turn, should encourage a consideration of these issues when selecting stimuli for an experiment.

Appendix B

eDom: Software and Norms for 544 English Homonyms

A central challenge in designing experiments to investigate semantic ambiguity is identifying an appropriate set of ambiguous words to serve as stimuli. One important factor to consider in this process is the relative interpretation frequency, or dominance, of each individual interpretation (Armstrong & Plaut, 2011; Frazier & Rayner, 1990; Klepousniotou & Baum, 2007; Swinney, 1979; Twilley et al., 1994). Dominance plays a key role in the processing of an ambiguous word because it influences the predictability of encountering each interpretation of the word. Prior to the presentation of a context that biases the selection of a particular interpretation, words with relatively balanced meaning frequencies and that lack a clearly dominant interpretation are least predictive of which interpretation should be activated (e.g., the <enclosure> and the <composite> meanings of COMPOUND occur approximately equally often). However, as the meaning frequencies become unbalanced and one meaning is clearly dominant, it is possible, in principle, to predict which meaning to activate with a high degree of accuracy (e.g., the <lease> meaning of RENT occurs more frequently than the <opening> meaning). As a result,

Note. A version of the work reported in this Appendix was also reported in Armstrong, Tokowicz, and Plaut (2012)

differences in dominance may critically modulate the processing of an ambiguous word by altering the representations that are activated and maintained in neutral contexts (Frazier & Rayner, 1990; Seidenberg et al., 1982; Swinney, 1979; Williams, 1992). When interpretation frequencies are balanced, both interpretations may partially be activated to the same extent. In contrast, when interpretation frequencies are unbalanced and one interpretation is clearly dominant, the dominant interpretation may be strongly activated and the activation of the subordinate interpretations may be substantially reduced (Seidenberg et al., 1982; Swinney, 1979). At the extremes, closely balanced ambiguous words may thus serve as the ideal items for use in many experiments investigating the effects of ambiguity, whereas strongly unbalanced ambiguous words may be virtually indistinguishable from the unambiguous words with which they are typically contrasted (Armstrong & Plaut, 2011; Hino et al., 2006; Klepousniotou & Baum, 2007; Rodd et al., 2002).

The difference between balanced and unbalanced ambiguous words is most pronounced in the case of homonyms, for which there is general agreement that the semantic overlap between interpretations is minimal (e.g., <dog>/<tree> BARK; Armstrong & Plaut, 2008; Frazier & Rayner, 1990; Hino et al., 2006; Klein & Murphy, 2001, 2002; Rodd et al., 2002, 2004; Rubenstein et al., 1970; Klepousniotou et al., 2008; Hino et al., 2010). This contrasts with polysemes, for which a single written and spoken form are associated with multiple *related* interpretations, which may reduce the degree to which each individual meaning may be differentially activated (Armstrong & Plaut, 2008, 2011; Beretta et al., 2005; Frazier & Rayner, 1990; Klepousniotou et al., 2008; Pylkkänen et al., 2006; Rodd et al., 2002, but see Hino et al., 2010, 2006; Klein & Murphy, 2001, 2002, for dissenting views). Consequently, assessing the dominance of homonyms is particularly important and failures to control for this factor have been proposed as an explanation for the weak and inconsistent effects of homonymy in many studies (Armstrong & Plaut, 2011).

The goals of the present work were threefold: 1) to develop an efficient and reliable technique for estimating the meaning frequencies of homonyms based primarily on ratings of dic-

tionary definitions, which avoids several problems with classic methods for assessing meaning frequency, 2) to collect normative data for a pool of words suitable for use in future investigations that could also be used to examine the characteristics of the norms and the reliability of the method, and 3) to demonstrate the predictive validity and utility of the norms by analyzing the results of lexical decision experiments reported previously by Armstrong and Plaut (2011) and as part of the English Lexicon Project (Balota et al., 2007). To this end, open-source software was devekioed to automate the norming process and aid in the collection of norms for 544 homonyms that have been made available for use and extension by other researchers. This appendix presents a brief overview of this work. Additional details, as well as the normative data and software are available in the online user manual located at `http://edom.cnbc.cmu.edu`.

B.1 Issues with Existing Norming Methods

One popular method for estimating dominance is via the classification of the free associates generated for a given homonym on the basis of the meaning of the word to which they are related (Geis & Winograd, 1974; Gilhooly & Logie, 1980a, 1980b; Gorfein, Viviani, & Leddo, 1982; Kausler & Kollasch, 1970; Mirman et al., 2010; Nelson, McEvoy, Walling, & Wheeler, 1980; Twilley et al., 1994), or related methods of classifying generated definitions (Warren, Bresnick, & Green, 1977), generated sentences (Wollen, Cox, Coahran, Shea, & Kirby, 1980), or sentence completions (Yates, 1978). These methods involve two steps: 1) participants are provided with an ambiguous word (e.g., BANK) and generate an associate (or other similar response; e.g., MONEY), and 2) a separate group of raters classify these responses on the basis of their intuitions regarding to which meaning they are related (e.g., the <financial> vs. the <edge of a river> meanings of BANK). Across a number of studies, this method has been shown to generate a fairly consistent measure of dominance (for a review, see Twilley et al., 1994).

There are, however, several issues with these methods. At a theoretical level, researchers must make the assumption that responses in the free association task are generated in direct proportion

to the relative frequency of each of the word's meanings. Although there may be some surface validity to this claim, correlations between norms generated via the free association technique, although high amongst other studies using the same methods, generally decrease when other similar techniques such as classification of generated sentences or definitions are used (Twilley et al., 1994). This suggests that there is a non-trivial task-specific component in the ratings generated by classifying free associates.

Bridging between the theoretical and methodological levels, raters often encounter difficulty in agreeing on which meaning each free associate should be linked with, if any—the average overlap in rater classifications was between only 65% and 75% in Twilley et al. (1994). Participants often produce associates that are not strongly semantically related to either interpretation of a homonym, which makes consistent classifications across raters difficult. This is true even under the assumption that participants and raters have identical semantic representations. For instance, if associations are weak, random noise in each raters' classification process may prevent consistent classification across raters. However, this assumption may not be valid because there may be differences in both the quantity and the type of discourse to which the raters and the participants have been exposed to, which in turn may cause each group to develop somewhat different semantic representations. At the very least, the low agreement across raters questions the efficacy of using data from free association tasks to generate relative meaning frequency ratings given the low information content of each associate/rating. At worst, it suggests that other types of non-semantic associations may influence response generation to a substantial degree. An examination of responses from free association tasks supports the latter conclusion: responses often consist of synonyms (e.g., COP $\Rightarrow$ OFFICER), antonyms (e.g., HOT $\Rightarrow$ COLD), category co-ordinates (e.g., ROBIN $\Rightarrow$ SPARROW), completions that form compound words (e.g., WRIST $\Rightarrow$ WATCH), and other associates that are not purely semantic in nature. These results invalidate the simple assumption that responses are semantic associates of the target and are generated as a function of the target's relative meaning frequencies. Clearly, a more refined

theory of how responses are generated in free association tasks is needed to improve the relative meaning frequency estimates that can be extracted from free association norms. It is therefore, worth developing a norming methodology that avoids these issues.

Another issue with the classification of free associates is how raters agree on the initial set of word meanings into which each response should be classified. To increase the consistency of this process, researchers often classify the associates into the meanings of the word listed in a dictionary (e.g., Mirman et al., 2010; Twilley et al., 1994). However, no evidence has been provided to show that dictionary definitions are sufficiently similar to the mental representations of word meanings in the target populations to be suitable for this task. For example, dictionary definitions may both fail to include representations of vernacular meanings and contain meanings based on the etymology of the word that are no longer in common usage, which may hinder the extraction of accurate meaning frequency estimates.

Finally, these techniques are resource intensive in several respects. The average number of participants that must generate associates is typically quite large in these studies, with many studies collecting ratings from well over 100 participants for only about 100 homonyms (Twilley et al., 1994). Additionally, each participant's responses must be classified individually by one, or preferably several, raters. With each study typically involving over 100 words, each rater must therefore classify over 10,000 observations. The end result of these resource demands is that large-scale norming studies are rare. Given continued concerns that a word's relative meaning frequencies can differ substantially across different participant populations and can change over time given the fluid nature of language (Swinney, 1979; Twilley et al., 1994), an alternative method that makes such norming studies more tractable is clearly desirable.

B.2 Dictionary Definition Rating Experiment

The present study investigated the reliability, validity, and efficiency of norms based on explicit relative meaning frequency ratings of dictionary definitions (supplemented by participant-

generated definitions) in a set of homonyms. From a theoretical standpoint, this method is a more direct assay of meaning frequency because it avoids the need to make assumptions about how the associates were generated. Further, it avoids the methodological drawbacks of investing considerable resources in having raters classify participants' responses (often according to the definitions listed in a dictionary) and the inconsistent classifications that often arise during this process across raters. Finally, the directness of this method (one participant ⇒ one dominance rating vs. many participants ⇒ classification of many responses ⇒ one dominance rating) should also lead to more rapid convergence on stable norms.

Methods

Participants. A total of 64 (24 male) native English speakers aged 18 or above and enrolled in psychology courses at the University of Pittsburgh participated in the experiment in exchange for course credit. There was no explicit screening to exclude participants with language disorders, although none were spontaneously reported by the participants.

Stimuli. Homonyms and a small number of homographs (i.e., words with a single orthographic form but two phonological forms associated with two different meanings, e.g., <turn>/<storm> WIND) that would be ideally suited for standard semantic ambiguity experiments were selected for norming based on the standard parameters of several variables (e.g., Armstrong & Plaut, 2011; Rodd et al., 2002). Specifically, these consisted of all words between 3 and 10 letters in length, with a log_{10}(SUBTL word frequency) between 1 and 100 (Brysbaert & New, 2009)[1], with sense counts in wordNet (Fellbaum, 1998), phoneme and syllable counts in N-Watch (Davis, 2005), and with two or more unrelated meanings in the Wordsmyth online dictionary, which has been employed in past semantic ambiguity studies (e.g., Armstrong & Plaut, 2011; Azuma & Van Orden, 1997; Rodd et al., 2002, and which is available online at wordsmyth.net, Parks et

[1]This upper bound also avoids the need to control for an additional quadratic frequency component due to a ceiling effect (Balota et al., 2004; Brysbaert & New, 2009).

al., 1998). Based on the classification scheme used in the online version of Wordsmyth, separate web page entries denoted unrelated meanings of a word, whereas separate definitions on a single page denoted distinct but related senses of the word. Further, the order in which each meaning appears reflects the rank-ordered frequency of that meaning according to the Wordsmyth lexographers. This coarse meaning/sense classification correlates with estimates of homonymy and polysemy (Azuma & Van Orden, 1997; Rodd et al., 2002). A total of 585 words satisfied these constraints—576 homonyms and 9 homographs, which were collapsed in with the homonyms in all of the remaining analyses—of which 483 had two meanings, 84 had three meanings, 15 had four meanings, and 3 had five meanings. A parsing script was used to extract the definitions for each of these words from Wordsmyth.[2] Given the large number of words that were normed, as well as the screening criteria employed for selecting these words, the results of the present norming study are broadly representative of the population of words that could appear in a typical experiment that employs homonymous stimuli.

Procedure. Due to the size of the word set, it was not feasible to have each participant rate each word in a single session. Instead, each participant rated a random sample of approximately one quarter of the full set, or 146 words, resulting in approximately 16 ratings per word in total. To ensure that each word was seen by an equal number of raters, these samples were generated by randomly sampling from the population without replacement until the population was exhausted, at which point the population was reset and the process was repeated. The randomization script which accomplish this is included with the *eDom* software.

Before beginning the experiment, participants were instructed that they would be estimating, as a percentage, how often a particular meaning of a presented word was implied when they encounter that word, and were given an example of a balanced and an unbalanced homonym. They were told that the dictionary definitions of the words were listed to remind them of the

[2]Five words were excluded because the automated parser did not correctly extract their definitions from Wordsmyth or there was a duplicate entry in the online dictionary.

meanings associated with the word and were instructed to read over the definitions to determine which meaning should be estimated. They were told that their estimates, however, should be based on their own personal experience, and the number, order, and length of the definitions should not directly impact their judgements. Additionally, they were instructed to list up to two additional meanings of a given word if they knew of meanings that were not included in the presented definitions. If additional meanings were listed, their relative meaning frequency ratings were to include ratings for those extra meanings. If they did not know a word at all, they were to respond "don't know". They were also instructed to try to be accurate in their ratings without spending too much time thinking about how to rate a particular word. Participants were prompted to ask any questions they had to the experimenter prior to beginning the experiment. The full instructions used in the experiment are available in the online manual.

A custom application called *eDom* was created to present the words and their definitions to participants. The full details of this software, as well as the source code and standalone binaries for several operating systems, are available to researchers via the online user manual. In brief, participants were presented with a 3 ×2 array in which each cell (left-to-right, top-to-bottom) contained all of the definitions of the related senses associated with a distinct meaning (see Figure B.1). The order in which the definitions associated with each meaning were presented was randomized so as to avoid presenting dominant and subordinate meanings based on their order of entry in Wordsmyth. Additionally, two of the cells were editable and shaded yellow and were available for participants to list other definitions of the words that they knew (only one such cell was available for the three stimuli with five meanings). Below each meaning was a field into which participants could enter their estimate of the meaning frequency, as a percentage of the time that meaning was implied when the word was encountered (default value = 0). New definitions listed by participants were required to have non-zero percentages, and the sum of all the percentages had to be 100. Once participants were done rating a word they pressed a "done rating" button and a new word was presented. Alternatively, if participants did not know any of

the meanings of a word they could press a “don’t know” button. Self-paced breaks were available after every 24 words. The entire experiment required approximately 50 minutes to complete.

Results

Removing data for unknown words. On average, participants responded that they did not know a word on 3% of trials. In total, there were 32 words for which at least approximately 20% of respondants (three participants) responded “don’t know” — these responses corresponded to 2.3% of the total responses. These words were excluded from further analysis along with all other trials for which a “don’t know” response was made. This left 553 words in the trimmed set, of which 544 were homonyms.

Participant supplementation of dictionary definitions. Participants listed additional meanings for the presented words on 2.6% of trials. These responses were distributed approximately equally across each block of the experiment. Twenty words had additional definitions given by at least 20% of respondants (three participants); the additional definitions for these words accounted for approximately 34% of all the participant-generated definitions. Of these words, one was part of the set that was removed due to exceeding the “unknown word” criterion, as detailed above. The first author manually reviewed each of the new definitions participants listed for these 20 words and noted the existence of a consistent new meaning each time that at least 20% of respondants’ definitions related to the same new meaning. In total, there were 11 words for which a consistent new meaning was identified and the majority of these definitions were related to vernacular meanings of the words (e.g., <muscular> BUFF). A list of these extra meanings is available in the online manual. The average relative meaning frequency of the participant-generated definitions was 42%, and only 24% these ratings had relative meaning frequencies greater than 50%. Taken together, the relative paucity of participant-generated definitions, combined with the moderate convergence of participants’ definitions on specific meanings, suggests that dictionary definitions are relatively successful at exhaustively capturing the meanings of most words. Ad-

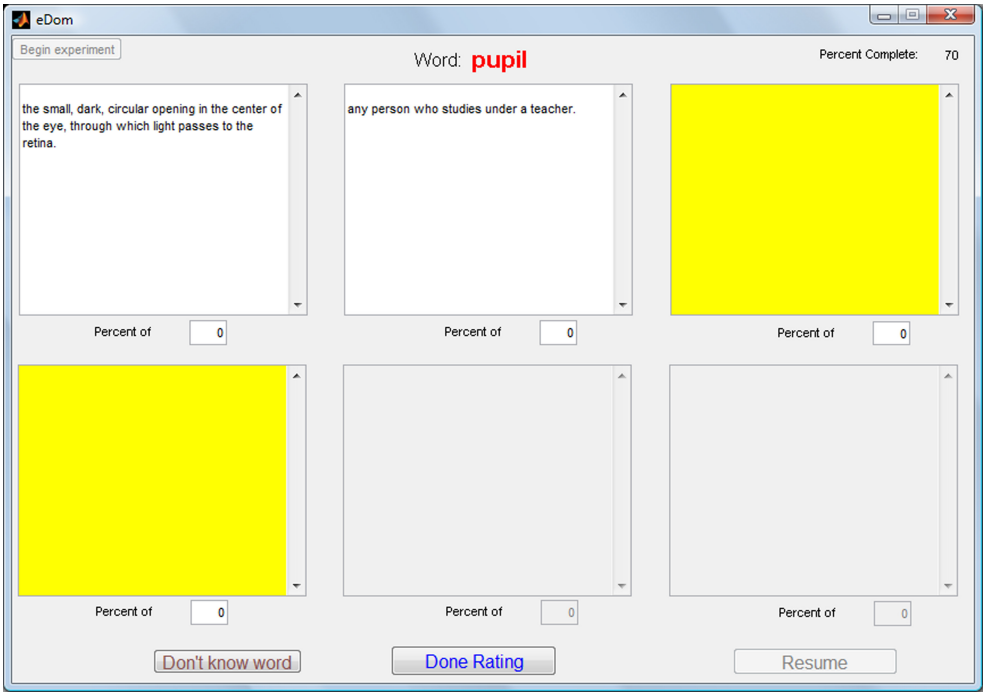


Figure B.1: Screen capture of the *eDom* norming software during the norming of the word PUPIL.

ditionally, in cases in which the dictionary definitions did not represent a common meaning of a word, participants supplemented the dictionary definitions with their own definitions.

Comparison of the meanings listed in the dictionary to those that were familiar to the participants. On average, the sum of the relative meaning frequencies across all of the dictionary definitions accounted for 99% of the sum of the relative meaning frequencies across all of the word's dictionary and participant supplemented definitions. This indicates that dictionary definitions come very close to constituting an exhaustive list of the meanings of homonyms. To determine whether some of these meanings were listed in the dictionary purely for etymological reasons and are not generally encountered by the participants, the average relative frequency ratings for the dictionary definitions were rank ordered and the summed rating across the first through second, third, fourth, and fifth dictionary meanings was calculated, where applicable. Summing across the two most frequent meanings accounted for the clear majority of encoun-

ters with a word (80% of encounters for words with four or five meanings, 90% of encounters for words with three meanings, and 99% of encounters for words with two meanings). These results imply that a number of the meanings associated with the words are either unknown to participants or are known but virtually never encountered. They also indicate that the bulk of encounters are captured by only the first and second meanings of each word, which allows for considerable simplification of the metric used to assess the dominance of a homonym (Twilley et al., 1994). On the basis of this observation, the characteristics of the normed words were examined using a very simple measure of dominance—the largest meaning frequency associated with each word, denoted as β, for *biggest*. Using the approximation that the sum of the relative meaning frequency ratings for the first two meanings captures all of the relative frequency data for all of the word's meanings, this measure is effectively a difference score between the first and second most frequent meanings.[3]

In addition to calculating the overall degree to which dictionary definitions span participants' representations of a word's meanings, additional tests examined whether the rank ordering of each definition of a word in Wordsmyth agreed with the rank ordering of participants' meaning frequency estimates. To do so, a score of 1 was assigned each time the rank ordering of the two most frequent meanings in each participant's ratings corresponded to the same rank ordering of the dictionary definitions, and 0 otherwise. These ratings were then collapsed across participants the average ratings were tested to determine whether the resulting scores tended to agree with the participant ratings. This was accomplished via a single-sample t-test with a null hypothesis corresponding to chance agreement (50%). The dictionary and participant ratings agreed 81% of the time on average, which was significantly higher than chance, $t(552) = 27, SE = 0.01, p <$

[3]This evidence not withstanding, the correlation between β and other more complex measures of dominance, such as the *B* measure introduced by Twilley et al. (1994) based on information-content theory and an alternative measure, *D*, which was developed based on a standardized difference between the largest and second largest meaning frequency $[(largest - secondLargest)/largest]$ were nevertheless very high (all *r*s $<= .94$). Similar results are therefore expected if using these other measures.

.0001.

Characteristics of the norms and implications for studies that employ homonymous stimuli.

Reliability. To evaluate the reliability of the ratings, the mean β for each word in the norms was computed and then each participant's ratings of a subset of these words were correlated with the mean ratings. Reasonable consistency between individual participants the mean ratings was observed ($r = .70; SE = 0.01; \text{range} = .41 - .85$). This consistency is comparable to that reported in a similar task in which participants rated the age of aquisition of each meaning of an ambiguous word (Khanna & Cortese, 2011). This provides initial support for the reliability of these norms and of their relative invariance to the particular set of words that form the context in which participants generate ratings.

Additional analyses examined whether averaging across the data from 16 participants per word, when each participant rated a different subset of the entire set of words, was sufficient to obtain stable estimates of β. To do so, the average β for each homonym obtained in the present norming study was compared to the β data collected in another another study using the same norming method (Armstrong & Plaut, 2011). In that study, 50 participants normed each of the 200 words with multiple meanings used in the experiment reported therein, of which 195 overlapped with those in the set of words from the present study. Thus, there were over three times as many observations (50 vs. 16) in the norms collected as part of the Armstrong and Plaut study. The correlation between the mean β values across these two sets was nevertheless very high ($r = .95$). Additionally, when the β ratings from the 16-participant data set were used to predict the β ratings from the 50-participant data set in a linear regression, the coefficient for the intercept was near zero ($b = 5.8; SE = 1.8$) and the coefficient for the 16-participant β was near 1 ($b = 0.95; SE = 0.02$). This confirms that the raw values from each norming experiment, and not simply the linear relationship between these values, were highly similar. Taken together, these

results indicate that stable ratings were obtained in the present study in which 16 participants rated each stimulus, and that the present norming procedure converges on reliable norms more rapidly and with fewer observations than other norming methodologies. Consequently, there is little to be gained by collecting data from more participants or requiring participants to rate the entire set of words. This contrasts with the original arguments Twilley et al. (1994) used to motivate their large-scale free association norming study.

Additionally, the reliability of the β measure was assessed as a function of the value of β by plotting the standard error of β as a function of β for each word[4] (see Figure B.2). Contrary to past research suggesting that the most balanced words are associated with the least stable ratings, (Swinney, 1979; Twilley et al., 1994), moderate variability was observed for β values between 40 and 60, high variability was observed between 60 and 90, and rapidly decreasing variability was observed between 90 and 100. The lack of a reduction in reliability for words with β values less than about 90 also provides support for the β measure as an alternative to more complex dominance metrics. For instance, these results do not support using a metric based on information-content theory such as the uncertainty measure U derived by Twilley et al. (1994), which loses sensitivity as words become more balanced.

Distribution of largest relative meaning frequency. Examining the distribution of β values provides insight into the expected characteristics of experimental words that have not controlled for relative meaning frequency during item selection. A histogram of the β scores for all of the normed homonyms is presented in Figure B.3. This figure shows that these data are moderately left-skewed and that the majority of the homonyms (~65%) have largest relative meaning frequencies in excess of 75%. Given that several studies have defined a homonym as having

[4] A second plot not included in the Appendix established that the relationship depicted in Figure B.2 was not due to a confound with several other variables. This was accomplished by plotting 'residual β', for which the correlated contributions from all of the variables listed in the 'Correlation with other variables' section had been removed. The resulting plot produced a qualitatively similar result to that presented here.

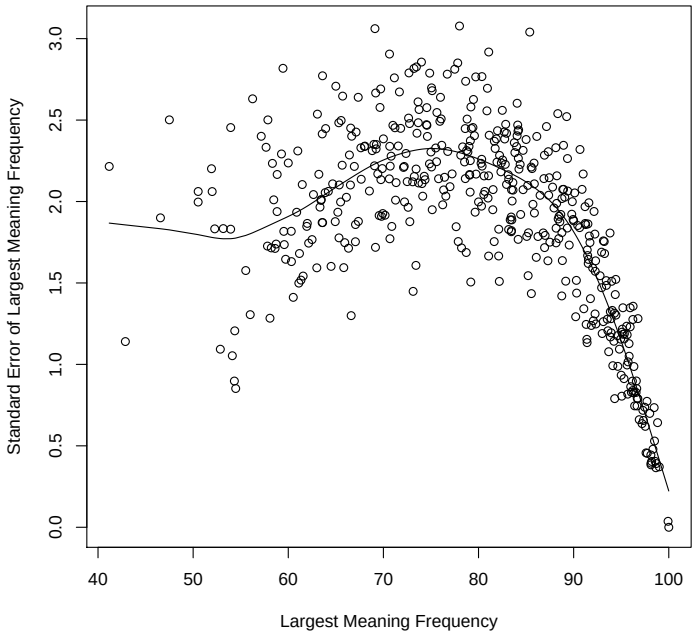


Figure B.2: Standard error of the largest meaning frequency estimates for each normed word as a function of each word's largest meaning frequency. The solid line depicts the mean of the standard error for all words with the same largest meaning frequency, fit with a cubic smoothing spline.

relatively balanced meaning frequencies if the largest relative meaning frequency was less than 75%, or an even more conservative value (Armstrong & Plaut, 2011; Klepousniotou & Baum, 2007; Mirman et al., 2010; Swinney, 1979), this indicates that a random sample of English words would fail to satisfy this constraint, given the relatively exhaustive coverage of the set of homonyms investigated in the present study. This finding emphasizes both the need to constrain item selection based on relative meaning frequency *a priori*, or otherwise address the effects of this variable—for instance, by including it as a covariate when analyzing the data—and the utility of the eDom norms to this end. This distribution may also help explain the small magnitude of the homonymy disadvantage that has been reported in several studies (Armstrong & Plaut, 2008, 2011; Beretta et al., 2005; Rodd et al., 2002). In many of these experiments, no consideration of dominance was used in selecting words and, as a result, over 50% of the words in those experiments had largest relative meaning frequenices above 75% (Armstrong & Plaut, 2011). Consequently, the weak or non-existent homonymy effects in these experiments may be due, at least in part, to the fact that the majority of words were closer to being unambiguous than they were to being optimally ambiguous.

Correlation with other variables. Whether β was significantly related to other common variables that are controlled for or experimentally manipulated in studies involving homonymous stimuli was assessed via correlation analyses. These correlations were significant ($p < .05$) when the correlation coefficient was greater than .10, given that there were 553 observations entered into each statistical test. Marginal effects ($p < .10$) are also indicated in the discussion below.

At the semantic level, significant correlations were observed with the number of unrelated meanings ($r = -.32$) and with the number of related senses ($r = -.28$), as well as with the number of interpretations[5] counts in wordNet ($r = -.27$). The correlations with these measures emphasize the need to carefully consider how to design studies that can detect strong main effects

[5]Typically these are refered to as 'sense counts', but the term 'interpretation' is used here to emphasize that these data are an aggregate measure of both the number of unrelated meanings and of related senses.

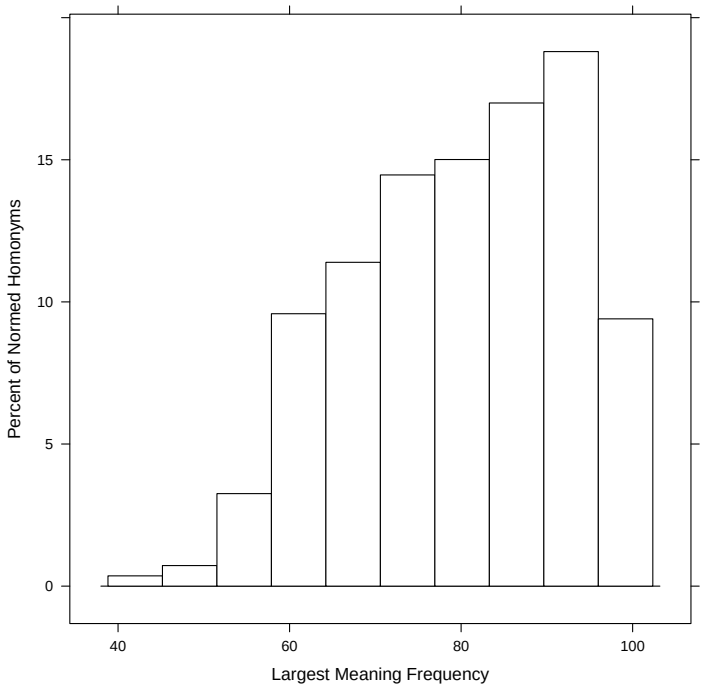


Figure B.3: Histogram of the distribution of the largest relative meaning frequency value for all of the normed words.

of relative meaning frequency, given that the number of meanings and number of senses associated with a word have been linked with processing disadvantages and advantages, respectively (Armstrong & Plaut, 2008, 2011; Klepousniotou & Baum, 2007; Rodd et al., 2002).

At a grammatical level, correlations with β and the number of distinct interpretations falling into particular grammatical classes, as listed in Wordsmyth and collapsed across meanings and senses, were observed for verbs ($r = -.26$) and nouns ($r = -.16$). No significant correlations were detected for adjectives ($r = -.07$) or adverbs ($r = -.05$). These data reveal that grammatical class is another important property to consider when using homonymous stimuli to study word and discourse comprehension, particularly with respect to nouns and verbs. It also opens the possibility for the different ambiguity effects that have been reported across grammatical classes to be the result of a confound with relative meaning frequency (Frazier & Rayner, 1990; Frisson & Pickering, 1999; Pickering & Frisson, 2001), although Mirman et al. (2010) observed different effects for noun-noun versus noun-verb homonyms even when controlling for a measure of relative meaning frequency.

At a lexical level, there were significant but weak correlations with both raw and log-transformed word frequency (rs $= -.11$), which agree with several past studies showing that such a relationship, if it exists, is weak (for discussion, see Twilley et al., 1994). The correlation with orthographic Levenshtein distance (Yarkoni et al., 2008) was also significant ($r = .14$), and the correlation with word length (in number of letters) was marginally significant ($r = .10$). No significant correlations were detected with sublexical measures of number of phonemes ($r = .04$), number of syllables ($r = .08$), or positional bigram frequency ($r = .09$), which supports the assumption that sublexical and semantic representations are independent (Armstrong & Plaut, 2008; Plaut, 1997).

Overall, these significant correlations point to the importance of carefully controlling or otherwise addressing these potential confounds when designing studies that investigate the effects of homonymy, relative meaning frequency, and many of the other factors listed above. The large

number of significant correlations also suggests that automated methods of eliminating these confounds (e.g., Armstrong, Watson, & Plaut, 2012) may be particularly valuable in such undertakings.

Comparison to relative frequency estimates derived from the classification of free associates. To evaluate the degree to which the definition-rating method of assessing dominance taps the same underlying representations as norms based on the classification of free associates, β was correlated with data from a large-scale norming study conducted using the latter methodology (Twilley et al., 1994). The Twilley et al. study reported a measure refered to as U, which consisted of a nonlinear transformation of the relative meaning frequencies for each meaning. However, the difference between expressing a word's relative meaning frequencies either as β or as U is very small for words with two meanings. This was determined by correlating these two measures across 21 equally-spaced data points which spanned the range of possible values of β/U for words with two meanings (i.e., β values in the range [.5, 1]). This yielded an r of .94. On the basis of this result and the fact that the bulk of the words in the present study were reported as essentially having only two meanings by the participants, it is therefore reasonable to assume that differences observed when comparing β from the present study to norms to U from Twilley et al. are primarily due to task differences. After adjusting for differences in the interpretation of the sign of the slope of each measure—β scores increase as a homonym's meanings become less balanced, whereas the opposite is true for U—the two measures correlated only weakly ($r = .27$). This result indicates that although these two measures tap a common underlying variable, for the most part they measure unique variance. Qualitatively, this is in line with the decrease in correlation between measures of relative meaning frequency observed across free associate classification, definition writing, and sentence generation (see Twilley et al.). Nevertheless, the correlations between the Twilley et al. data and the data obtained using those alternative methodologies were considerably stronger (smallest $r = .72$). Additional work is needed to more fully understand the cause of these differences. The comparison of the predic-

tive validity of norms generated by each of these methods, presented next, provides guidance to this end.

Predictive validity. The predictive validity of the largest relative meaning frequency measure was assessed in regression analyses involving data from two different lexical decision studies, described below.

Analysis of the data reported by Armstrong and Plaut (2011). Several previous studies have investigated how homonymy influences word comprehension using the visual lexical decision task (e.g., Armstrong & Plaut, 2008; Azuma & Van Orden, 1997; Klepousniotou & Baum, 2007; Rodd et al., 2002). These studies suggest that once the task is made difficult—for instance, by matching the orthographic characteristics of the words and nonwords to reduce the informativeness of orthography for selecting a response—homonyms show a processing disadvantage relative to unambiguous controls (see Armstrong & Plaut, 2008, for a mechanistic account of this effect). However, this processing disadvantage has typically been weak and has often failed to reach statistical significance ($p < .05$). Armstrong and Plaut (2011) hypothesized that this was in part attributable to failing to control for relative meaning frequency—an issue that was discovered in *post hoc* norming of the words used in a previous study. Additionally, they hypothesized that these effects would be stronger if even more difficult versions of the lexical decision task were employed.

To address these issues, Armstrong and Plaut (2011) examined whether task difficulty influences the magnitude of the homonomy effect by testing this effect with the same word stimuli in four conditions formed by crossing nonword difficulty (easy vs. difficult; the analyses reported here exclude the preliminary findings reported in that paper concerning a third nonword condition in which pseudohomophones were presented) and stimulus contrast (i.e., presenting words as white-on-black vs. grey-on-black; high contrast vs. low contrast). Because the high contrast conditions yielded no significant homonymy effects in that study, the re-analysis reported here

are based on only the low contrast conditions. TSpecifically, these analyses focus on stimuli that were labelled as (relatively) 'unambiguous controls' and 'homonyms' in that paper. Because it is relevant later, it is also worth noting that these stimuli were further constrained such that there were few related senses associated with each of the meanings of the homonyms and unambiguous controls. This somewhat relaxed definition of what constitutes a 'homonym' was necessary because 'pure' homonyms which have only distinct, unrelated interpretations are extremely rare — there are only 38 such words in the norms reported here, and only 8 of these words have β scores below 75%. Critically, however, the unambiguous controls and homonyms were matched to have the same total number of related senses summed across all of their interpretations. Consequently, these conditions differ only in terms of whether these interpretations are clustered as a single set of related interpretations or are spread across two distinct interpretations. Differences between these conditions can therefore be attributed to whether the words were homonymous or not. Given that relative frequency norms were not available for a large set of stimuli when Armstrong and Plaut selected words for use in their experiment, they did not control for this factor when selecting their experimental stimuli. Instead, they made the *a priori* decision to norm the homonyms they selected using the method reported in this appendix while collecting data in the lexical decision task, and to control for relative meaning frequency during later analyses.

The homonyms and unambiguous controls from each of the the easy and difficult lexical decision experiments were compared in three different analyses. In the first analysis, Number of meaningS (one vs. many) was entered as a factor and analyzed the full data set. In the second analysis, Number of Meanings was also entered as a factor but the homonym condition was restricted to include only the data for 14 well-balanced homonyms. These homonyms had a mean β of 59% and all of the homonyms had a β less than or equal to 65%. This operational definition of a 'balanced' homonym is similar to that employed by Klepousniotou and Baum (2007). Essentially the same results were also obtained using an upper-bound of 75% as in Mirman et al. (2010), which allowed for approximately double the number of homonyms to

be entered into the analysis. In the third analysis, instead of entering Number of Meanings as a factor, β was entered as a continuous variable and analyzed the full data set (unambiguous controls were assigned a β of 100). These are refered to as the 'full-factor', 'restricted-factor', and 'full-regression' analyses. All of the analyses were conducted using mixed-effect regression and *p*-values for the statistical tests were calculated via Monte Carlo simulations (Baayen et al., 2008). In addition to the Number of Meanings factor or β variables, these analyses also included log_{10}(SUBTL word frequency), word length (number of letters), orthographic Levenshtein distance, number of phonemes, number of syllables, residual familiarity[6], trial rank, and the lexicalty, accuracy, and latency of the previous trial (Baayen & Milin, 2010) as fixed effects, and participants and words as random effects. All of these fixed and random effects succeeded in predicting significant amounts of variance in at least a subset of the analyses conducted by Armstrong and Plaut (2011), and additional predictors such as imageability and positional bigram frequency were not included because they did not predict significant amounts of variance in Armstrong and Plaut's analyses. Only correct responses were included in the latency analyses, which were measured in milliseconds. For ease of interpretation, in the remainder of the appendix the slopes for the different coefficients have been standardized such that a positive slope in accuracy and a negative slope in latency always indicates a homonymy disadvantage.

In the full-factor analyses, no significant effects of homonymy were observed ([easy nonwords] accuracy: b = 0.0001, SE = 0.0083, p = .99, n = 9933; latency: b = -6.8, SE = 4.1, p = .09, n = 9129. [hard nonwords] accuracy: b = 0.0115, SE = 0.0096, p = .23, n = 10439; latency: b = 1.2, SE = 4.4, p = .78, n = 9562). However, a significant or marginal homonymy disadvantage was observed in all of the latency analyses that included some consideration of the effects of relative meaning frequency, with the exception of the 'hard' nonword condition in

[6]Residual familiarity was calculated by regressing out the effects of the meaning and frequency factors from the raw familiarity scores. As noted in Armstrong and Plaut (2011), these two measures correlated very strongly (r = .98) and so, for consistency with the analyses in previous work, residual familiarity was used in all of the present analyses.

the resticted-factor analysis ([easy nonwords] restricted-factor: b = -16.7, SE = 8.3, p = .04, n = 5301; full-regression: b = -0.35, SE = 0.16, p = .02, n = 9129. [hard nonwords] restricted-factor: b = -7.6, SE = 9.0, p = .37, n = 5597; full-regression: b = -0.29, SE = 0.18 p = .09, n = 9562). Several effects approached significance in the accuracy analyses, as well ([easy non-words] restricted-factor: b = 0.0326, SE = 0.0168 p = .05, n = 5818; full-regression: b = 0.0002, SE = 0.0003, p = .55, n = 9933. [hard nonwords] restricted-factor: b = 0.0257, SE = 0.0162, p = .12, n = 6110; full-regression: b = 0.0006, SE = 0.0004, p = .10, n = 10439). Further discussion of these results is defered until after presenting the analyses of a second set of lexical decision data.

Using β and U to predict lexical decision performance in the English Lexicon Project. To further assess the predictive validity of β, and to directly compare the predictiveness of β to that of U, the following analyses examined how well each measure predicted performance in the lexical decision task[7] conducted as part of the English Lexicon project (Balota et al., 2007). In this lexical decision task, words were presented at full contrast and uses much easier nonwords than Armstrong and Plaut (2011); together, these differences in procedure may alter the effect of homonymy. Indeed, data from the English Lexicon Project (although not from the specific set of words analyzed here) were reported recently to show a homonymy advantage, even when smaller-scale lexical decision experiments have failed to reach significance on these comparisons (Hargreaves et al., 2011). The exact cause of these opposing effects of homonymy is an unresolved issue in the literature and beyond the scope of the present work (for several related accounts, see Hargreaves et al., 2011; Hino & Lupker, 1996; Hino et al., 2006; Kawamoto, 1993). Here the focus is on whether different measures of relative meaning frequency can contribute to understanding these effects by increasing the magnitude and reliability of the homonymy advantage reported using this data set.

[7]The naming data were not used because semantic ambiguity effects are typically weak or non-existent in those data (Borowsky & Masson, 1996).

The English lexicon project contained data for 551 of the normed words in the present data set and for 211 words that appeared in both the present data set and the Twilley et al. (1994) norms. Given the substantial differences in the number of available words, two sets of analyses were completed. The first included only the words for which the U measure was available so as to compare both β and U on equal footing, and the second included the full data set to maximize the generalizbility and statistical power of the analysis. For brevity, the different analyses involving β are refered to via different subscripts on the β coefficients: f for the *full* data set and u for the data set for which the U measure was available.

First, β and U were used to predict accuracy and latency using simple regression. These analyses showed that β was a significant or marginal predictor of both measures of performance ([accuracy] β_f: $b = -0.0006, SE = 0.0003$, p = .01; β_u: $b = -0.0004, SE = 0.0002$, p = .05. [latency (in ms)] β_f: $b = 0.4, SE = 0.2$, p = .07; β_u: $b = 0.6, SE = 0.3$, p = .05), however, U was not ([accuracy] $b = 0.001, SE = 0.006$, p = .77. [latency] $b = -13, SE = 8$, p = .10). Specifically, β predicted higher accuracies and shorter latencies for more balanced homonyms. Notwithstanding, statistical tests between the slopes associated with normalized variants of each of these metrics were not significantly different, although there was a weak trend in the case of accuracy ([accuracy] d = 0.030, SE = 0.019, $t(418)$ = 1.6, p = .11. [latency] d = 5.96, SE = 34.2, $t(418)$ = 0.24, p = .81).

Next, robustness of these findings was examined in multiple regression analyses that also included several additional independent variables. The results reported next were produced by analyses that included length (in letters), log_{10}(SUBTL word frequency), orthographic Levenshtein distance, positional bigram frequency, number of phonemes, number of syllables, number of senses, number of verb interpretations, and number of noun interpretations as predictors. Simultaneous multiple regression was employed in which no interactions were allowed amongst these variables. Separate analyses were conducted for β and U to avoid collinearity issues. Only the statistics related to the relative frequency measurements are reported. In these analyses, nei-

ther variable predicted significant variance in either the accuracy or the latency data, although the effect of U was marginal when predicting latency ([accuracy] β_f: $b = -0.0003, SE = -0.0002$, p = .16; β_u: $b = -0.0003, SE = 0.0002$, p = .23; U: $b = 0.0005, SE = 0.005$, p = .93. [latency] β_f: $b = -0.06, SE = 0.19$, p = .76; β_u: $b = 0.4, SE = 0.3$, p = .18; U: $b = -13, SE = 7$, p = .08). Finally, these regression analyses were repeated on a restricted set of the data that contained only the homonyms with fewer than 10 related senses associated with each of their meanings. This restriction limits the analyses to a similar subset of homonyms to those reported in the context of the analysis of the Armstrong and Plaut (2011) data, and which the aforementioned paper suggested may reveal stronger effects of homonymy. There were 397 such homonyms in the full data set and 110 in the data set for which U data was available. Only the analyses of β related to the set of words for which U data were available reached significance ([accuracy] β_f: $b = -0.0005, SE = 0.0003$, p = .11; β_u: $b = -0.0009, SE = 0.0004$, p = .04; U: $b = -0.005, SE = 0.01$, p = .66. [latency] β_f: $b = 0.05, SE = 0.3$, p = .24; β_u: $b = 0.95, SE = 0.46$, p = .04; U: $b = -10, SE = 11$, p = .39). The difference between the slopes associated with normalized variants of each of these metrics was marginal in the accuracy analysis (d = 0.050, $SE = 0.029$, $t(216)$ = 1.75, p = .08) but was not significant in the latency analysis (d = 24.0, SE = 30.2, $t(216)$ = 0.79, p = .43). Consequently, although the simple regression analyses weakly suggested that β was a significant predictor of a homonymy advantage (whereas U was not), the subsequent multiple regression analyses suggest that this effect is at best extremely weak and is limited to homonyms with few related senses.

Discussion. In the analyses of the Armstrong and Plaut (2011) data, relative meaning frequency clearly altered the observed effects. Specifically, restricting the analyses to homonyms for which relative frequencies were balanced or including relative meaning frequency as a continuous variable generally allowed for the detection of a significant homonymy disadvantage. In contrast, the analyses of the data from the English lexicon project provide only very weak support for a general homonymy advantage in those data, and even that effect is restricted to when β, as

opposed to U, is entered as a predictor. Of course, the weakness of many of these effects—which is attributable at least in part to not explicitly considering relative meaning frequency when selecting the experimental words–precludes drawing strong conclusions from these results. Nevertheless, these results support two tentative conclusions. First, relative meaning frequency should be considered when selecting experimental stimuli to maximize what appear to be, at best, weak effects of homonymy. Second, the presence of a homonymy advantage is particularly suspect and it may be preferable to investigate the effects of homonymy in more difficult variants of the lexical decision task. However, additional experimental work using a range of tasks and carefully controlled sets of words will be needed to substantiate and generalize these conclusions. Automated means of selecting optimal stimuli (e.g., Armstrong, Watson, & Plaut, 2012) may be particularly helpful in these endeavors to maximize the magnitude of the apparently weak effects of homonymy in visual lexical decision. Furthermore, the effects of homonymy should be examined in other tasks to establish the robustness of the present findings. Auditory lexical decision in particular may be worthy of further study, because there are some data that suggest that the effects of homonymy are stronger in that task (Klepousniotou & Baum, 2007; Mirman et al., 2010; Rodd et al., 2002).

B.3 Conclusion

The present work outlines a method of assessing the relative meaning frequencies of each meaning of a homonym based on dictionary definitions. This approach offers several theoretical and methodological advantages over standard norming approaches based on the classification of free associates. Normative data collected with the *eDom* software program for a large set of homonyms provides empirical evidence that supports the reliability and validity of this approach. Although additional work is needed to better understand the unique aspects of meaning frequency that this method and other methods tap, as well as the effects of homonymy more generally, the present results support the conclusion that this norming method is to be prefered

over standard methods based on free association.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for boltzmann machines. *Cognitive Science*, *9*(1), 147–169.

Akiyama, Y., Yamashita, A., Kajiura, M., & Aiso, H. (1989). Combinatorial optimization with gaussian machines. *Proceedings of the International Joint Conference on Neural Networks*, 533–540.

Anderson, J. A. (1991). Why, having so many neurons, do we have so few thoughts? In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 477–507). Hillsdale, NJ: Erlbaum.

Armstrong, B. C. (2007). *Comprehending Ambiguous Words: Computational and Empirical Investigations.* Masters thesis, National Archives of Canada.

Armstrong, B. C., Joordens, S., & Plaut, D. C. (2009). Yoked criteria shifts in decision system adaptation: Computational and behavioral investigations. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the* 31st *Annual Conference of the Cognitive Science Society* (pp. 2130–2135). Austin, TX: Cognitive Science Society.

Armstrong, B. C., LeBoutillier, J. C., & Petit, T. L. (2012). Ultrastructural synaptic changes associated with neurofibromatosis type 1: A quantitative analysis of hippocampal region ca1 in a nf1+/- mouse model. *Synapse*, *66*(3), 246–255.

Armstrong, B. C., & Plaut, D. C. (2008). Settling dynamics in distributed networks explain task differences in semantic ambiguity effects: Computational and behavioral evidence. In B. C. Love, K. McRae, & V. M. Sloutsky (Eds.), *Proceedings of the* 30th *Annual Con-*

ference of the Cognitive Science Society (pp. 273–278). Austin, TX: Cognitive Science Society.

Armstrong, B. C., & Plaut, D. C. (2011). Inducing homonymy effects via stimulus quality and (not) nonword difficulty: Implications for models of semantic ambiguity and word recognition. In L. Carlson, C. Hölscher, & T. Shipley (Eds.), *Proceedings of the* 33rd *Annual Conference of the Cognitive Science Society* (pp. 2223–2228). Austin, TX: Cognitive Science Society.

Armstrong, B. C., Tokowicz, N., & Plaut, D. C. (2012). edom: Norming software and relative meaning frequencies for 544 english homonyms. *Behavior Research Methods*.

Armstrong, B. C., Watson, C. E., & Plaut, D. C. (2012). SOS: An algorithm and software for the stochastic optimization of stimuli. *Behavior Research Methods*.

Azuma, T., & Van Orden, G. C. (1997). Why SAFE is better than FAST: The relatedness of a word's meanings affects lexical decision times. *Journal of Memory and Language*, *36*(4), 484–504.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baayen, R. H., & Milin, P. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*(2), 12–28.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (release 2)[cd-rom]. *Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania [Distributor]*.

Bainbridge, J. V., Lewandowsky, S., & Kirsner, K. (1993). Context effects in repetition priming are sense effects. *Memory and Cognition*, *21*, 619–619.

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access?: The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology*, *10*(3), 340–357.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General*, *133*(2), 283–316.

Balota, D. A., & Duchek, J. M. (1991). Semantic priming effects, lexical repetition effects, and contextual disambiguation effects in healthy aged individuals and individuals with senile dementia of the alzheimer type. *Brain and Language*, *40*(2), 181 - 201.

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., ... Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–449.

Becker, C. A. (1980). Semantic context effects in visual word recognition: An analysis of semantic strategies. *Memory & Cognition*, *8*(6), 493–512.

Beretta, A., Fiorentino, R., & Poeppel, D. (2005). The effects of homonymy and polysemy on lexical access: an MEG study. *Cognitive Brain Research*, *24*(1), 57–65.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, *19*, 2767–2796.

Bogacz, R., Brown, E., Moehlis, J., Holmes, P., & Cohen, J. (2006). The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, *113*(4), 700.

Bogacz, R., Usher, M., Zhang, J., & McClelland, J. L. (2007). Extending a biologically inspired model of choice: Multi-alternatives, nonlinearity and value-based multidimensional choice. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1485), 1655–1670.

Booth, J. R., Harasaki, Y., & Burman, D. D. (2006). Development of lexical and sentence level context effects for dominant and subordinate word meanings of homonyms. *Journal of Psycholinguistic Research*, *35*(6), 531–554.

Borowsky, R., & Besner, D. (1993). Visual word recognition: A multistage activation model.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *19*(4), 813–840.

Borowsky, R., & Besner, D. (2006). Parallel Distributed Processing and Lexical-Semantic Effects in Visual Word Recognition: Are a Few Stages Necessary? *Psychological Review*, *113*(1), 181–193.

Borowsky, R., & Masson, M. E. J. (1996). Semantic ambiguity effects in word identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(1), 63–85.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624.

Botvinick, M. M., Nystrom, L., Fissell, K., Carter, C., & Cohen, J. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature*, *402*(6758), 179–181.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*(4), 433-436.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990.

Bunn, E. M., Tyler, L. K., & Moss, H. E. (1998). Category-specific semantic deficits: The role of familiarity and property type reexamined. *Neuropsychology*, *12*(3), 367–379.

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, *12*(4), 335–359.

Clopper, C. J., & Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413.

Coltheart, M. (1981). The MRC psycholinguistic database. *The Quarterly Journal of Experimental Psychology*, *33*(4), 497–505.

Coltheart, M., Davelaar, E., Jonasson, J. T., & Besner, D. (1977). Access to the internal lexicon. *Attention and Performance*, 535–555.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2000). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*,

108(1), 204–256.

Connor, S., Williams, P. T. J., Armstrong, B., Petit, T. L., Ivanco, T. L., & Weeks, A. C. W. (2006). Long-term potentiation is associated with changes in synaptic ultrastructure in the rat neocortex. *Synapse*, *59*(6), 378–382.

Conrad, C. (1974). Context effects in sentence comprehension: A study of the subjective lexicon. *Memory & Cognition*, *2*(1), 130–138.

Cortese, M. J., & Khanna, M. M. (2007). Age of acquisition predicts naming and lexical-decision performance above and beyond 22 other predictor variables: An analysis of 2,342 words. *Quarterly Journal of Experimental Psychology*, *60*(8), 1072–1082.

Cree, G. S., & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello(and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163.

Cutler, A. (1981). Making up materials is a confounded nuisance, or: Will we be able to run any psycholinguistic experiments at all in 1990? *Cognition*, *10*(1-3), 65–70.

Dagerman, K. S., MacDonald, M. C., & Harm, M. W. (2006). Aging and the use of context in ambiguity resolution: Complex changes from simple slowing. *Cognitive Science*, *30*(2), 311–345.

Davis, C. J. (2005). N-WATCH: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods, Instruments, & Computers*, *37*(1), 65–70.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446.

Espinoza-Varas, B., & Watson, C. S. (1994). Effects of decision criterion on response latencies of binary decisions. *Attention, Perception, & Psychophysics*, *55*(2), 190–203.

Faust, M. E., Balota, D. A., Duchek, J. M., Gernsbacher, M. A., & Smith, S. (1997). Inhibitory control during sentence comprehension in individuals with dementia of the Alzheimer type. *Brain and Language*, *57*(2), 225–253.

Federmeier, K. D., & Laszlo, S. (2009). Time for meaning: Electrophysiology provides insights into the dynamics of representation and processing in semantic memory. *Psychology of Learning and Motivation*, *51*, 1–44.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. The MIT press.

Forster, K. I., & Bednall, E. S. (1976). Terminating and exhaustive search in lexical access. *Memory & Cognition*, 53–61.

Frazier, L., & Rayner, K. (1990). Taking on semantic commitments: Processing multiple meanings vs. multiple senses. *Journal of Memory and Language*, *29*(2), 181–200.

Frisson, S., & Pickering, M. J. (1999). The processing of metonymy: Evidence from eye movements. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(6), 1366–1383.

Geis, M. F., & Winograd, E. (1974). Norms of semantic encoding variability for fifty homographs. *Bulletin of the Psychonomic Society*, *3*(6), 429–431.

Gernsbacher, M. A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness, and polysemy. *Journal of Experimental Psychology: General*, *113*(2), 256–281.

Gernsbacher, M. A., & Faust, M. E. (1991). The mechanism of suppression: A component of general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(2), 245–262.

Gibson, F. P., Fichman, M., & Plaut, D. C. (1997). Learning in dynamic decision tasks: Computational model and empirical evidence. *Organizational Behavior and Human Decision Processes*, *71*, 1–36.

Gilhooly, K., & Logie, R. (1980a). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods*, *12*(4), 395-427.

Gilhooly, K., & Logie, R. (1980b). Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods*, *12*(4),

428–450.

Gilhooly, K., & Logie, R. (1982). Word age-of-acquisition and lexical decision making. *Acta Psychologia*, *50*, 21–34.

Gomez, P., Ratcliff, R., & Perea, M. (2007). A model of the go/no-go task. *Journal of Experimental Psychology: General*, *136*(3), 389.

Gorfein, D. S., Viviani, J. M., & Leddo, J. (1982). Norms as a tool for the study of homography. *Memory & Cognition*, *10*(5), 503–509.

Gottlob, L. R., Goldinger, S. D., Stone, G. O., & Van Orden, G. C. (1999). Reading homographs: Orthographic, phonologic, and semantic dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, *99*(S3), 561–574.

Grainger, J., & Jacobs, A. M. (1996). Orthographic processing in visual word recognition: A multiple read-out model. *Psychological Review*, *103*(3), 518–565.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211.

Hanes, D. P., & Schall, J. D. (1996). Neural control of voluntary movement initiation. *Science*, *274*(5286), 427–430.

Hargreaves, I. S., Pexman, P. M., Pittman, D. J., & Goodyear, B. G. (2011). Ambiguous words recruit the left inferior frontal gyrus in absence of a behavioral effect. *Experimental Psychology*, *58*(1), 19–30.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491–528.

Healy, M. J. R. (1968). Multiple regression with a singular matrix. *Journal of the Royal Statistical Society: Series C*, *17*(2), 110–117.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the stroop task. *Psychological Bulletin*, *109*(2), 340.

Hino, Y., Kusunose, Y., & Lupker, S. J. (2010). The relatedness-of-meaning effect for ambiguous

words in lexical-decision tasks: when does relatedness matter? *Canadian Journal of Experimental Psychology*, *64*(3), 180–196.

Hino, Y., & Lupker, S. J. (1996). Effects of polysemy in lexical decision and naming: An alternative to lexical access accounts. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(6), 1331–1356.

Hino, Y., Lupker, S. J., & Pexman, P. M. (2002). Ambiguity and synonymy effects in lexical decision, naming, and semantic categorization tasks: Interactions between orthography, phonology, and semantics. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(4), 686–713.

Hino, Y., Pexman, P. M., & Lupker, S. J. (2006). Ambiguity and relatedness effects in semantic tasks: Are they due to semantic coding? *Journal of Memory and Language*, *55*(2), 247–273.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*(1-3), 185–234.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1: Foundations* (pp. 282–317). Cambridge, MA: MIT Press.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*(1), 74.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 598–615.

Hogaboam, T. W., & Perfetti, C. A. (1975). Lexical ambiguity and sentence comprehension.

Journal of Verbal Learning and Verbal Behavior, *14*(3), 265–274.

Houghton, G., & Tipper, S. P. (1996). Inhibitory mechanisms of neural and cognitive control: Applications to selective attention and sequential action. *Brain and Cognition*, *30*(1), 20–43.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*(1), 106–154.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks*, *1*(4), 295–307.

Jastrzembski, J. E. (1981). Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon. *Cognitive Psychology*, *13*(2), 278–305.

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review*, *17*(5), 662–672.

Joordens, S., & Besner, D. (1994). When banking on meaning is not (yet) money in the bank: Explorations in connectionist modeling. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(5), 1051–1062.

Joordens, S., Piercey, C. D., & Azarbehi, R. (2009). Modeling performance at the trial level within a diffusion framework: A simple yet powerful method for increasing efficiency via error detection and correction. *Canadian Journal of Experimental Psychology*, *63*(2), 81–93.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of Neural Science* (4th ed.). McGraw-Hill New York.

Kausler, D. H., & Kollasch, S. F. (1970). Word associations to homographs. *Journal of Verbal Learning and Verbal Behavior*, *9*(4), 444–449.

Kawamoto, A. H. (1993). Nonlinear dynamics in the resolution of lexical ambiguity: A parallel distributed processing account. *Journal of Memory and Language*, *32*(4), 474–516.

Kawamoto, A. H., Farrar, W. T., & Kello, C. T. (1994). When two meanings are better than

one: Modeling the ambiguity advantage using a recurrent distributed network. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1233–1247.

Kellas, G., Ferraro, F. R., & Simpson, G. B. (1988). Lexical ambiguity and the timecourse of attentional allocation in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(4), 601–609.

Khanna, M. M., & Boland, J. E. (2010). Children's use of language context in lexical ambiguity resolution. *The Quarterly Journal of Experimental Psychology*, *63*(1), 160–193.

Khanna, M. M., & Cortese, M. J. (2011). Age of acquisition estimates for 1,208 ambiguous and polysemous words. *Behavior Research Methods*, *43*(1), 89–96.

Kiefer, J., & Wolfowitz, J. (1952). Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, *23*(3), 462–466.

Kintsch, W., & Mross, E. F. (1985). Context effects in word identification. *Journal of Memory and Language*, *24*(3), 336–349.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, *220*(4598), 671–680.

Klein, D. E., & Murphy, G. L. (2001). The representation of polysemous words. *Journal of Memory and Language*, *45*(2), 259–282.

Klein, D. E., & Murphy, G. L. (2002). Paper has been my ruin: Conceptual relations of polysemous senses. *Journal of Memory and Language*, *47*(4), 548–570.

Klepousniotou, E. (2002). The processing of lexical ambiguity: Homonymy and polysemy in the mental lexicon. *Brain and Language*, *81*(1-3), 205–223.

Klepousniotou, E., & Baum, S. R. (2007). Disambiguating the ambiguity advantage effect in word recognition: An advantage for polysemous but not homonymous words. *Journal of Neurolinguistics*, *20*(1), 1–24.

Klepousniotou, E., Titone, D., & Romero, C. (2008). Making sense of word senses: The comprehension of polysemy depends on sense overlap. *Journal of Experimental Psychology:*

Learning, Memory, and Cognition, *34*(6), 1534–1543.

Krajbich, I., & Rangel, A. (2011). Multialternative drift-diffusion model predicts the relationship between visual fixations and choice in value-based decisions. *Proceedings of the National Academy of Sciences*, *108*(33), 13852–13857.

Kučera, H., & Francis, W. N. (1967). *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Laszlo, S., & Federmeier, K. D. (2008). Minding the PS, queues, and PXQs: Uniformity of semantic processing across multiple stimulus types. *Psychophysiology*, *45*(3), 458–466.

Laszlo, S., & Federmeier, K. D. (2011). The N400 as a snapshot of interactive processing: Evidence from regression analyses of orthographic neighbor and lexical associate effects. *Psychophysiology*, *48*(2), 176–186.

Laszlo, S., & Plaut, D. C. (2011). Simulating event-related potential reading data in a neurally plausible parallel distributed processing model. In *Proceedings of the* 33rd *Annual Conference of the Cognitive Science Society.* Mahwah, NJ: Lawrence Erlbaum Associates.

Laughlin, S., & Sejnowski, T. (2003). Communication in neuronal networks. *Science*, *301*(5641), 1870–1874.

Lehrer, A. (1990). Polysemy, conventionality, and the structure of the lexicon. *Cognitive Linguistics*, *1*(2), 207–246.

Link, S. W. (1975). The relative judgement theory of two choice response time. *Journal of Mathematical Psychology*, *12*(1), 114–135.

Link, S. W., & Heath, R. (1975). A sequential theory of psychological discrimination. *Psychometrika*, *40*, 77–105.

Locker, L., Simpson, G. B., & Yates, M. (2003). Semantic neighborhood effects on the recognition of ambiguous words. *Memory & Cognition*, *31*(4), 505.

MacDonald, A. W., Cohen, J. D., Stenger, V. A., & Carter, C. S. (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, *288*(5472), 1835–1838.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Lawrence Erlbaum.

Martin, C., Vu, H., Kellas, G., & Metcalf, K. (1999). Strength of discourse context as a determinant of the subordinate bias effect. *The Quarterly Journal of Experimental Psychology Section A*, *52*(4), 813–839.

Mason, R. A., & Just, M. A. (2007). Lexical ambiguity in sentence comprehension. *Brain Research*, *1146*, 115–127.

Masson, M. E. J., & Borowsky, R. (1995). Unsettling questions about semantic ambiguity in connectionist models: Comment on Joordens and Besner (1994). *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*, 509–509.

Masson, M. E. J., & Freedman, L. (1990). Fluent identification of repeated words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*(3), 355–373.

McClelland, J. L. (1993). The GRAIN model: A framework for modeling the dynamics of information processing. *Attention and Performance XIV: Synergies in Experimental Psychology, Artifical Intelligence, and Cognitive Neuroscience*, 655–688.

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

McClelland, J. L., & Elman, J. L. (1986). The trace model of speech perception. *Cognitive Psychology*, *18*(1), 1–86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complemen-

tary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*(3), 419–457.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of the effect of context in perception: Part 1. *Psychological Review*, *88*, 375–407.

McClelland, J. L., St John, M., & Taraban, R. (1989). Sentence processing: A parallel distributed processing approach. *Language and Cognitive Processes: Special issue on Parsing and Interpretation*, *4*(3/4), 287–335.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

Mehler, J., Segui, J., & Carey, P. (1978). Tails of words: Monitoring ambiguity. *Journal of Verbal Learning and Verbal Behavior*, *17*(1), 29–35.

Mewhort, D. J. K. (2005). A comparison of the randomization test with the F test when error is skewed. *Behavior Research Methods*, *37*(3), 426–435.

Millis, M. L., & Button, S. B. (1989). The effect of polysemy on lexical decision time: Now you see it, now you don't. *Memory & Cognition*, *17*(2), 141–147.

Mirman, D., Strauss, T. J., Dixon, J. A., & Magnuson, J. S. (2010). Effect of representational distance between meanings on recognition of ambiguous spoken words. *Cognitive Science*, *34*(1), 161–173.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(1), 116.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with

symmetric diffusion networks. *Cognitive Science*, *17*(4), 463–496.

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*(4), 413–429.

Neely, J. H. (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G. Humphreys (Eds.), *Basic Processes in Reading: Visual Word Recognition.* Hillsdale, N.J.: Earlbaum.

Neill, W. T., Hilliard, D. V., & Cooper, E. A. (1988). The detection of lexical ambiguity: Evidence for context-sensitive parallel access. *Journal of Memory and Language*, *27*(3), 279–287.

Nelson, D. L., McEvoy, C. L., Walling, J. R., & Wheeler, J. W. (1980). The University of South Florida homograph norms. *Behavior Research Methods*, *12*(1), 16–37.

Newman, J. E., & Dell, G. S. (1978). The phonological nature of phoneme monitoring: A critique of some ambiguity studies. *Journal of Verbal Learning and Verbal Behavior*, *17*(3), 359–374.

Newman, R. L., & Joanisse, M. F. (2011). Modulation of brain regions involved in word recognition by homophonous stimuli: An fMRI study. *Brain Research*, *1367*, 250–264.

Nievas, F., & Marí-Beffa, P. (2002). Negative priming from the non-selected meaning of the homograph. *British Journal of Psychology*, *93*(1), 47–66.

Nunberg, G. (1979). The non-uniqueness of semantic solutions: Polysemy. *Linguistics and Philosophy*, *3*(2), 143–184.

Oden, G. C., & Spira, J. L. (1983). Influence of context on the activation and selection of ambiguous word senses. *The Quarterly Journal of Experimental Psychology Section A*, *35*(1), 51–64.

Onifer, W., & Swinney, D. A. (1981). Accessing lexical ambiguities during sentence comprehension: Effects of frequency of meaning and contextual bias. *Memory & Cognition*, *9*(3), 225–236.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cumberland, RI: The MIT Press.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery and meaningfulness values for 925 nouns. *Journal of Experimental Psychology*, *76*, 1–25.

Parks, R., Ray, J., & Bland, S. (1998). *Wordsmyth English Dictionary-thesaurus [Retrieved September 2008 from wordsmyth.net]* (Vol. 1).

Patterson, K., & Plaut, D. C. (2009). Shallow draughts intoxicate the brain: Lessons from cognitive science for cognitive neuropsychology. *Topics in Cognitive Science*, *1*(1), 39–58.

Paul, S. T., Kellas, G., Martin, M., & Clark, M. B. (1992). Influence of contextual features on the activation of ambiguous word meanings. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 703–717.

Peirce, J. W. (2007). Psychopy–psychophysics software in python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13.

Perea, M., Rosa, E., & Gómez, C. (2002). Is the go/no-go lexical decision task an alternative to the yes/no lexical decision task? *Memory & Cognition*, *30*(1), 34–45.

Pexman, P., Hargreaves, I. S., Siakaluk, P. D., Bodner, G. E., & Pope, J. (2008). There are many ways to be rich: Effects of three measures of semantic richness on visual word recognition. *Psychonomic Bulletin & Review*, *15*(1), 161–167.

Pexman, P., Hino, Y., & Lupker, S. J. (2004). Semantic ambiguity and the process of generating meaning from print. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(6), 1252–1270.

Pexman, P., & Lupker, S. J. (1999). Ambiguity and visual word recognition: Can feedback explain both homophone and polysemy effects? *Canadian Journal of Experimental Psychology*, *53*(4), 323–334.

Pexman, P., Lupker, S. J., & Hino, Y. (2002). The impact of feedback semantics in visual word

recognition: Number-of-features effects in lexical decision and naming tasks. *Psychonomic Bulletin & Review*, *9*(3), 542–549.

Pickering, M. J., & Frisson, S. (2001). Processing ambiguous verbs: Evidence from eye movements. *Learning & Memory*, *27*(2), 556–573.

Piercey, C. D., & Joordens, S. (2000). Turning an advantage into a disadvantage: Ambiguity effects in lexical decision versus reading tasks. *Memory & Cognition*, *28*(4), 657–666.

Pitt, M., & Samuel, A. (2006). Word length and lexical activation: Longer is better. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(5), 1120–1135.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the* 17th *Annual Conference of the Cognitive Science Society* (p. 37).

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Langugage and Cognitive Processes*, *12*(5/6), 765-805.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*(4), 786.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115.

Plaut, D. C., Nowlan, S., & Hinton, G. E. (1986). *Experiments on learning by back-propagation* (Tech. Rep. No. CMU-CS-86-126.) Pittsburgh, PA: Department of Computer Science, Carnegie-Mellon University.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500.

Purcell, B. A., Heitz, R. P., Cohen, J. Y., Schall, J. D., Logan, G. D., & Palmeri, T. J. (2010). Neurally constrained modeling of perceptual decision making. *Psychological Review*, *117*(4),

1113–1143.

Pylkkänen, L., Llinás, R., & Murphy, G. L. (2006). The representation of polysemy: MEG evidence. *Journal of Cognitive Neuroscience*, *18*(1), 97–109.

Pylkkänen, L., Stringfellow, A., & Marantz, A. (2002). Neuromagnetic evidence for the timing of lexical activation: An MEG component sensitive to phonotactic probability but not to neighborhood density. *Brain and Language*, *81*, 666–678.

R Development Core Team. (2011). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/` (ISBN 3-900051-07-0)

Raaijmakers, J. G. W., Schrijnemakers, J., & Gremmen, F. (1999). How to deal with "The language-as-fixed-effect fallacy": Common misconceptions and alternative solutions. *Journal of Memory and Language*, *41*, 416–426.

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *The Quarterly Journal of Experimental Psychology Section A*, *55*(4), 1339–1362.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*(2), 59–108.

Ratcliff, R. (1980). A note on modeling accumulation of information when the rate of accumulation changes over time. *Journal of Mathematical Psychology*, *21*(2), 178–184.

Ratcliff, R. (1988). Continuous versus discrete information processing: Modeling accumulation of partial information. *Psychological Review*, *95*(2), 238–255.

Ratcliff, R. (2002). A diffusion model account of response time and accuracy in a brightness discrimination task: Fitting real data and failing to fit fake but plausible data. *Psychonomic Bulletin & Review*, *9*(2), 278–291.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*(1), 159.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-

choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Rouder, J. N. (1998). Modeling response times for two-choice decisions. *Psychological Science*, *9*(5), 347–356.

Ratcliff, R., Thapar, A., Gomez, P., & McKoon, G. (2004). A diffusion model analysis of the effects of aging in the lexical-decision task. *Psychology and Aging*, *19*(2), 278.

Ratcliff, R., Thapar, A., & McKoon, G. (2003). A diffusion model analysis of the effects of aging on brightness discrimination. *Attention, Perception, & Psychophysics*, *65*(4), 523–535.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*(2), 261–300.

Reder, L. M. (1983). What kind of pitcher can a catcher fill? Effects of priming in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior*, *22*(2), 189–202.

Reichle, E. D., Pollatsek, A., Fisher, D. L., & Rayner, K. (1998). Toward a model of eye movement control in reading. *Psychological Review*, *105*(1), 125–157.

Rodd, J. M., Gaskell, G., & Marslen-Wilson, W. (2002). Making sense of semantic ambiguity: Semantic competition in lexical access. *Journal of Memory and Language*, *46*(2), 245–266.

Rodd, J. M., Gaskell, M. G., & Marslen-Wilson, W. D. (2004). Modelling the effects of semantic ambiguity in word recognition. *Cognitive Science*, *28*(1), 89–104.

Rodd, J. M., Johnsrude, I. S., & Davis, M. H. (2010). The role of domain-general frontal systems in language comprehension: Evidence from dual-task interference and semantic ambiguity. *Brain and Language*, *115*(3), 182–188.

Rubenstein, H., Garfield, L., & Millikan, J. A. (1970). Homographic entries in the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *9*(5), 487–494.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In *Backpropagation: Theory, architectures, and applications* (pp. 1–34). Hillsdale, New Jersey, USA: Lawrence Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagation errors. *Nature*, *323*, 533–536.

Rumelhart, D. E., & McClelland, J. L. (1986a). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition: Foundations (Volume 1)*. Cambridge, MA: The MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986b). *Parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Parallel distributed models of schemata and sequential thought processes. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 2: Psychological and Biological Models* (pp. 7–57). Cambridge, MA: MIT Press.

Salazar, R., Plastino, A. R., & Toral, R. (2000). Weakly nonextensive thermostatistics and the Ising model with long-range interactions. *European Physical Journal B*, *17*(4), 679–688.

Schilling, H. E. H., Rayner, K., & Chumbley, J. I. (1998). Comparing naming, lexical decision, and eye fixation times: Word frequency effects and individual differences. *Memory & Cognition*, *26*(6), 1270–1281.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime, Version 1.1 [Computer Software]*. Pittsburgh, PA: Psychology Software Tools.

Schneider, W., Eschman, A., & Zuccolotto, A. (2010). *E-prime, Version 2.0.8.90 [Computer Software]*. Pittsburgh, PA: Psychology Software Tools.

Schvaneveldt, R. W., Meyer, D. E., & Becker, C. A. (1976). Lexical ambiguity, semantic context, and visual word recognition. *Journal of Experimental Psychology*, *2*(2), 243–256.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. M., & Bienkowski, M. (1982). Automatic access of the meanings of ambiguous words in context: Some limitations of knowledge-based processing. *Cognitive Psychology*, *14*(4), 489–537.

Seidenberg, M. S., Waters, G. S., Sanders, M., & Langer, P. (1984). Pre-and postlexical loci of contextual effects on word recognition. *Memory & Cognition*, *12*(4), 315–328.

Sereno, S. C., Rayner, K., & Posner, M. I. (1998). Establishing a time-line of word recognition: Evidence from eye movements and event-related potentials. *NeuroReport*, *9*(10), 2195–2200.

Siakaluk, P. D., Buchanan, L., & Westbury, C. (2003). The effect of semantic distance in yes/no and go/no-go semantic categorization tasks. *Memory & Cognition*, *31*(1), 100–113.

Siakaluk, P. D., Pexman, P. M., Sears, C. R., & Owen, W. J. (2007). Multiple meanings are not necessarily a disadvantage in semantic processing: Evidence from homophone effects in semantic categorisation. *Language and Cognitive Processes*, *22*(3), 453–467.

Simpson, G. B. (1981). Meaning dominance and semantic context in the processing of lexical ambiguity. *Journal of Verbal Learning and Verbal Behavior*, *20*(1), 120–136.

Simpson, G. B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, *96*(2), 316–340.

Simpson, G. B. (1994). Context and the processing of ambiguous words. In Morton Ann Gersnbacher (Ed.), *Handbook of Psycholinguistics* (pp. 359–374). San Diego, CA: Academic Press.

Simpson, G. B., & Foster, M. R. (1986). Lexical ambiguity and children's word recognition. *Developmental Psychology*, *22*(2), 147–154.

Simpson, G. B., Lorsbach, T. C., & Whitehouse, D. (1983). Encoding and contextual components of word recognition in good and poor readers. *Journal of Experimental Child Psychology*, *35*(1), 161–171.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *Trends in*

Neurosciences, *27*(3), 161–168.

Snodgrass, J. G., & Vanderwart, M. (1980). A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(2), 174.

Stanovich, K. E. (2001). *How to Think Straight about Psychology (5th ed.).* New Jersey: Allyn and Bacon.

Sternberg, S. (1969). Memory-scanning: Mental processes revealed by reaction-time experiments. *American Scientist*, *57*(4), 421–457.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*, 251–260.

Swaab, T. Y., Brown, C., & Hagoort, P. (1998). Understanding ambiguous words in sentence contexts: Electrophysiological evidence for delayed contextual selection in broca's aphasia. *Neuropsychologia*, *36*(8), 737 - 761.

Swinney, D. A. (1979). Lexical access during sentence comprehension:(Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659.

Tabossi, P. (1988). Accessing lexical ambiguity in different types of sentential contexts. *Journal of Memory and Language*, *27*(3), 324–340.

Taler, V., Klepousniotou, E., & Phillips, N. A. (2009). Comprehension of lexical ambiguity in healthy aging, mild cognitive impairment, and mild alzheimer's disease. *Neuropsychologia*, *47*(5), 1332–1343.

Tipper, S. P., & Driver, J. (1988). Negative priming between pictures and words in a selective attention task: Evidence for semantic processing of ignored stimuli. *Memory & Cognition*, *16*(1), 64–70.

Tsetsos, K., Usher, M., & McClelland, J. (2011). Testing multi-alternative decision models with non-stationary evidence. *Frontiers in Neuroscience*, *5*.

Twilley, L. C., Dixon, P., Taylor, D., & Clark, K. (1994). University of Alberta norms of relative meaning frequency for 566 homographs. *Memory & Cognition*, *22*(1), 111–126.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*(3), 550–592.

Usher, M., & McClelland, J. L. (2004). Loss aversion and inhibition in dynamical models of multialternative choice. *Psychological Review*, *111*(3), 757–769.

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with speed-accuracy tradeoff. *Journal of Mathematical Psychology*, *46*(6), 704–715.

van Casteren, M., & Davis, M. (2007). Match: A program to assist in matching the conditions of factorial experiments. *Behavior Research Methods*, *39*(4), 973–978.

Van Petten, C., & Kutas, M. (1987). Ambiguous words in context: An event-related potential analysis of time the course of meaning activation. *Journal of Memory and Language*, *26*(2), 188–208.

Vickers, D., & Lee, M. (2000). Dynamic models of simple judgments: Ii. properties of a self-organizing pagan (parallel, adaptive, generalized accumulator network) model for multi-choice tasks. *Nonlinear Dynamics, Psychology, and Life Sciences*, *4*, 1–31.

Warren, R. E., Bresnick, J. H., & Green, J. P. (1977). Definitional dominance distributions for 20 English homographs. *Bulletin of the Psychonomic Society*, *10*(3), 229–231.

Watson, C. E. (2009). *Computational and behavioral studies of normal and impaired noun/verb processing*. Unpublished doctoral dissertation, Carnegie Mellon University.

Watson, C. E., & Chatterjee, A. (2012). A bilateral frontoparietal network underlies visuospatial analogical reasoning. *NeuroImage*, *59*, 2831–2838.

Wike, E. L., & Church, J. D. (1976). Comments on Clark's "The language-as-fixed-effect fallacy". *Journal of Verbal Learning and Verbal Behavior*, *15*(3), 249–255.

Williams, J. N. (1992). Processing polysemous words in context: Evidence for interrelated meanings. *Journal of Psycholinguistic Research*, *21*(3), 193–218.

Wollen, K. A., Cox, S. D., Coahran, M. M., Shea, D. S., & Kirby, R. F. (1980). Frequency of occurrence and concreteness ratings of homograph meanings. *Behavior Research Methods*,

12(1), 8–15.

Woollams, A. M., Lambon Ralph, M., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*(2), 316–339.

Yarkoni, T., Balota, D. A., & Yap, M. (2008). Moving beyond Coltheart's N: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*(5), 971-979.

Yates, J. (1978). Priming dominant and unusual senses of ambiguous words. *Memory & Cognition*, *6*(6), 636–643.