Blocking in semantic access task.

In the field of Cognitive Linguistics, there exists a certain controversy over semantic knowledge and meaning. Positions generally fall into one of two categories. Grounded, distributed, or embodied views posit that meaning or knowledge is represented in brain regions strongly implicated in perceptual and motor functions associated with the content of knowledge. On these views, retrieval or use of knowledge involves activating the relevant perceptual and motor regions; to understand “pink pillow” or to know that pillows are soft, one’s brain must recruit the visual and tactile sensory systems, very much in the same way as if the individual were seeing or feeling the pillow. Amodal, or ‘lexiconal’ views hold that semantic knowledge or meaning is held in a dedicated module, regardless of where that knowledge came from, and that retrieval involves matching some set of words to a set of facts about it. Both sides have degrees of strength. Stronger views of the latter will include a strongly modular theory of the brain, in which the brain can be said to be a bunch of theoretically separable pieces, each with its own discrete task. Stronger versions of the former will include a simulation theory of meaning, which states that to understand language about red pillows, one must recruit sensory portions of one’s brain, perform an offline simulation of the experience of red pillows, and generate meaning somewhere in between.

Inspired by a successful study performed by Kutas, Amsel, and Urbach that successfully demonstrated strong evidence for the embodied theory, we sought to recreate their results in a testing environment that strengthens the scope of the conclusion. Whereas previously semantic access was indexed by ERP components, we hoped to observe the same effects when semantic access was indexed by reaction time. Our initial study failed to obtain, and analysis of the stims and results concluded that a lack of statistical strength, and major sensitivity to outlying or variable data, were the reasons for the failure. I note the design of this initial experiment below. In a revised version of this initial experiment, the number of samples per subject was doubled, and so similarly was statistical strength. My experiment was designed to be a sister study of this revised experiment, differing from it in only one, key, experimental manipulation. This manipulation allows for evidence of the mechanism that causes the embodied effects in the initial ERP experiment, that we hope to see in the sister study.

Initial experimental design

Our hypothesis was that manipulating environmental factors that differentially affect the auditory and visual perceptual systems would in turn affect the time frame of access to semantic knowledge about properties in these modalities differentially. More specifically, we hypothesized that low contrast would increase response time for both auditory and visual content stimuli as compared to high contrast, but that the increase for visual stimuli would be significantly larger. This is because affecting contrast has, according to Amsel, Urbach, and Kutas, "been observed to modulate neural activity in several hierarchical visual regions."

We had subjects respond to sequential word pairs on screen. The first of any word pair was always a property; second words were always objects. Subjects were tasked with deciding if the first word was a valid or an invalid property of the second word, for every pair, and to signal their choice with a button press. Responses were recorded with keyboards, and were mapped to the a and l keys for every trial. Subjects were instructed to rest their hands over the base of the keyboard and area just off, so that only short, subtle movements of the index finger were ever required.

The entirety of the experiment featured a background grey screen with an RGB value of (121,121,121). For every data instance, subjects first observed a sole fixation point for 500 ms. This fixation point was the character "■," presented on screen with an RGB value of (126,126,126), in a 10 point font. The fixation point remained on screen throughout every instance and, because there was no visible change between instances, seemed to be continuously onscreen for every instance. Following this fixation period, the property word would appear for 300 milliseconds, followed by an interstimulus interval of 300 ms, and then an object word would appear for 300 ms. The ISI screen was identical to the fixation screen. Words appeared in 18 point helvetica font, and appeared centered just above the fixation square. Subjects had finite time of varying length to respond to any word pair, so that maintaining focus would be required throughout. This response interval ended at a random time between 2200 and 2400 ms after the object word disappeared.

Stimuli.

We used a list of word pairs, composed of one modifier and one noun each. Half of all word pairs described an auditory relationship (crying baby, blaring horn, screeching siren), and the other half described visual relationships (tall dwarf, shiny diamond, sparkling jewels). Half of either of these categories would describe semantically valid relationships (furry bear, transparent water), wherein the other half would describe invalid relationships (tall dwarf, crying crib, quiet roar). All stimuli were selected to remain within a certain narrow range of word frequency, pair-wise collocation frequency, and semantic similarity, as a counterbalancing measure for individual item difficulty. In addition, for each participant, we would randomly assign half of each of the four possible conditions to be presented in high contrast, with the others to be presented in low contrast. Low and high contrast in this case was just text with an RGB value of (126,126,126), and (139,139,139), respectively, presented against the grey background of (121,121,121).

Procedure

Subjects sat in a dark room, facing a computer screen with the grey background This background persisted throughout the whole of the experiment. The procedure had three phases, with only the last of the three involving any experimental stimuli or data logging. The first phase was a vision test. The word pairs presented in this phase were nonsensical (e.g. sluggish lightning), and uniformly presented in low contrast. Subjects were merely asked to read out loud the pairs as they appeared. Subjects who couldn't see a majority of the stimuli were disqualified from the study; however, the subject pool consistently scored perfect or close to perfect on the vision test. The most common mistake was over the first stimulus, which suggests that subjects initially underestimated the difficulty of the task, but adjusted readily.

Phase two was designed to familiarize subjects with the task, to practice and solidify the motor movements required for the experimental portion, and to allow us a chance to correct any behavior that could compromise the experiment. Here subjects received the full set of instructions. For every word pair on screen, they were asked to respond with either "valid" or "invalid" according to the semantic validity of the stimulus. Inputs were recorded with a keyboard, where the "L" and "A" keys were the available responses. Half of the subjects were assigned the key corresponding to their dominant hand to represent "valid," the other half were instructed to use the key on their non-dominant side. Subjects were observed as they

performed phase II, and were aware that data wasn't being logged. The task was identical to the experimental portion, except that it was shorter, and used its own, separate set of word pairs, which were again nonsensical.

Phase three was identical in basic structure to phase two. The experimenter would leave the room, and the subject would begin reading and judging word pairs. The experimental stimuli described above were used exhaustively. Subjects responded as above to approximately 34 word pairs, took an indefinite break ended by subject input, responded to an additional ~34 word pairs, took a second break, and then finished all the remaining stimuli to conclude the trial. The fixation square was on screen throughout the experiment, occluded only for break screens.

Analysis

We omitted any incorrectly answered trials, as well as all the invalid trials, correctly answered or otherwise. We then ran an ANOVA, taking the means of all four modality/contrast conditions. We subtracted means within modality, across contrast first, then subtracted those means between modality.

Results were not as promising as we'd hoped. We had a grand total of 104 experimental stimuli per subject. Only 52 were valid, and of those 52, we had only 13 of every possible contrast/modality pairing. This is a rather statistically weak design, and the ANOVA we ran made this clear. We did get the effect we were looking for, but there was a huge variance for the relevant parameters. Strength tests demonstrated that the result was statistically insignificant. The difference of the two differences (high.visual - low.visual) - (high.auditory - low.auditory) was positive as we'd hoped, but the relative weakness of our design necessitates a near uniform result to achieve statistical significance. This particular flaw in our design was changed for our revised studies. Notably, for both studies, we have doubled the stimuli number, by which we increased statistical strength.

Revised sister studies

My experiment was run near identically to the one just described, with identical stimulus parameters, procedure, instructions to the subject, etc. We dropped the counterbalancing across handedness to validity, because invalid trials are experimentally irrelevant, and no loss of effect results from having all subjects respond with their dominant hands. Stimuli number was approximately doubled. 24 new modifiers were

introduced, and of these 128 running modifiers, 79 appeared twice in the new design. These were designed such that half of these modifiers appeared in the same validity on both occasions, the other half appeared in different validities. This counter-balancing measure was included to prevent the development of any strategy associated with repeated modifiers, so that subjects were forced to rely on semantic knowledge in all cases. The break design was implemented just the same. However, there would be six blocks compared to the previous three, in line with an approximate doubling of stimuli.

The above would be true of both sister studies. The addition central to my own experiment is a blocking by contrast design. Each block was exclusively in a single contrast setting, and sequential blocks throughout an experiment would alternate in contrast. Half of all subjects observed a low-odd, high-even design, while the other half observed a high-odd, low-even design.

There are many theoretical advantages of this block design. First, and most importantly, is that it allows us to disambiguate the mechanisms underlying the observed results in the initial ERP experiment, as we hoped to see in the sister, non-blocked experiment. The possible results of this measure yields three alternative hypotheses. If the interaction effect in the blocked experiment obtains, then we can rule out a "switch-cost" explanation from our interpretation of the data. This result would be strongest, showing that, in general, if more visual area is dedicated to processing a stimulus, language comprehension about visual relationships or content is delayed, even across a significant portion of time and habituation to the stimuli features. We will call this result hypothesis 1. If the interaction effect in the blocked experiment obtains, but is in the opposite direction from the expected, then we observe an advantage for visual low contrast stimuli. This advantage would demonstrate that more recruitment of visual areas to process the stimuli differentially counteracts the main effects of contrast differentially for visual and auditory content language. This means that the visual region can, and does, simultaneously work for processing stimuli and simulating for language meaning. This rules out recruitment inhibition, and strongly suggests that switch-cost is a key feature of the results observed in the initial ERP experiment, hoped to be observed in the sister study. This result we'll call hypothesis 2. Finally, If our blocked interaction effect does not obtain, and the sister study interaction effect does, this suggests very strongly that dissonance or noise associated with switch-cost is what produced the

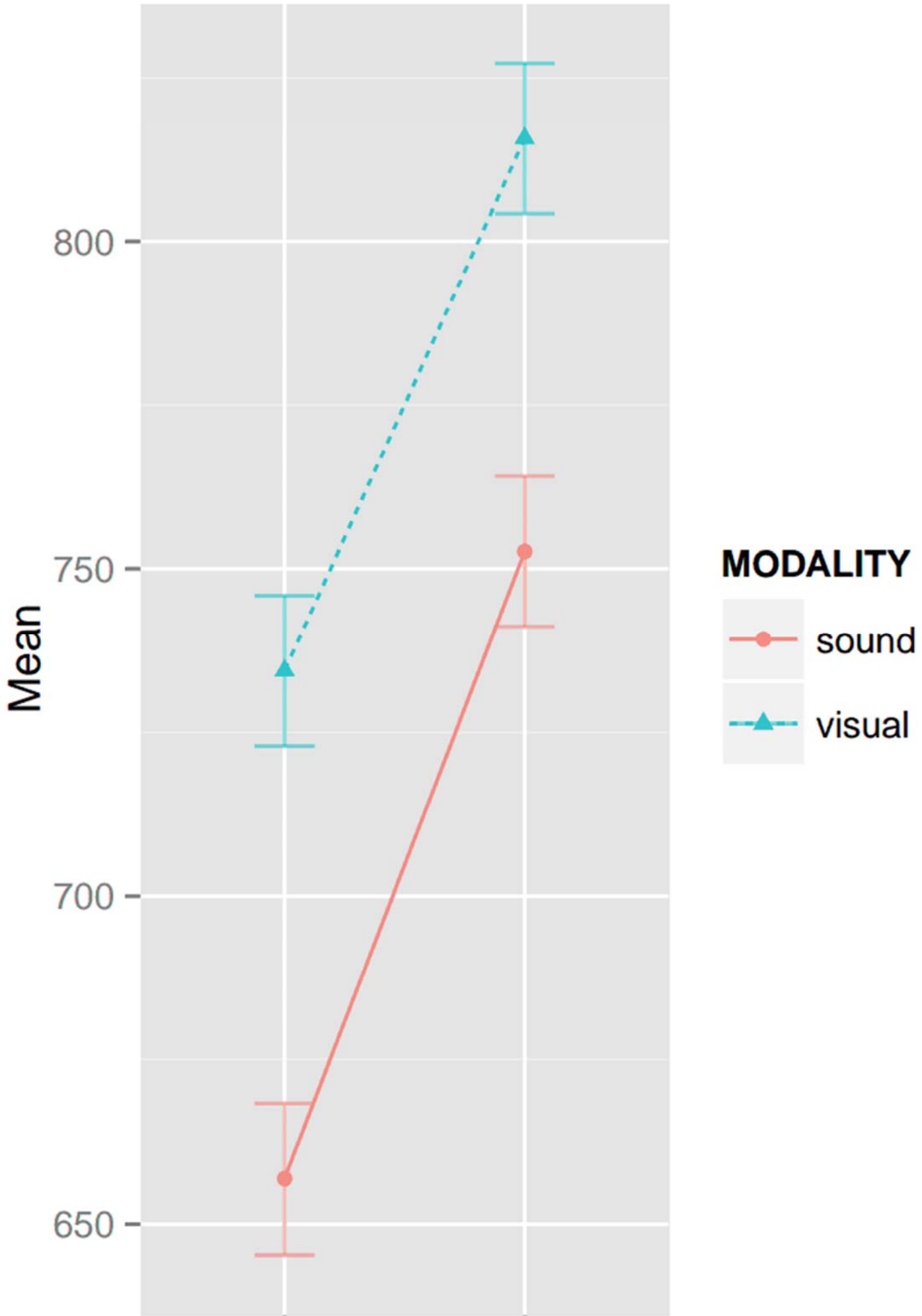
interaction effects in the other studies, as this switch-cost is the only feasible effect present in the obtaining studies, but absent in mine. We'll call this hypothesis 3.

Certain block effects can accumulate throughout a block design. If we observe that subjects consistently improve reaction time throughout a block, we can measure the interaction effect across stimulus presentation order within a block. Steady increase in interaction effect, in either direction, would be strong evidence for hypothesis 1 or 2, and would serve to further delineate the mechanisms of embodied effects on semantic access.

A final advantage to a block design is the possible leeway into future experiments. In block studies we can see the magnitude and direction of block effects on reaction time, independently of contrast manipulations. We can measure distance into a block where sustained modality is advantageous, disadvantageous, or neutral to performance on reaction time measures. This allows us to hone in on the time-scale of the relevant sensory computations, and this might be useful to extend our hypothesis. If, for example, we see something of a fatigue effect, where an initial advantage from sustained modality hits a threshold of negative return, we can infer that the relevant sensory area is fatiguing. This might work as evidence towards the stronger view of the sensory computation as an offline simulation, rather than just simple recruitment.

Block Design Results

We did not obtain a statistically significant interaction effect in the blocked experiment. We cleaned the data, removing all trials with a reaction time less than 200 ms or over 2000 ms. This was to omit outliers that either took too long or responded too quickly to be considered legitimate trials. We also dropped all subjects with an average accuracy less than 70%, which turned out to be just one subject averaging about 65%. We then averaged reaction time for all experimental condition combinations, auditory high, auditory low, visual high, and visual low. We saw a main effect of contrast that was very significant, as both auditory and visual modality trials dropped off significantly in low contrast. Our interaction effect did not obtain statistical significance; we observed a value of 2.3 ms (figure below). This would, assuming the success of the sister study, suggest heavily the choice of hypothesis 3.


800
750
700
650
Mean
MODALITY
sound
visual

Unfortunately, the non-blocked sister study had to be dropped due to a mistake in procedure formatting. As a result, we do not know if the same stimuli as used in the blocked experiment will create an interaction effect when presented in randomly sequenced contrast settings, as they did in the initial ERP study. Without the results of this experiment, we cannot reliably conclude with any of our main hypotheses, which require the results of the sister study as a control.

Analysis

There are three ways of interpreting the lack of a significant interaction effect in the blocked study. The first is that the initial ERP study was poorly conducted, and that the theory underlying our hypothesis is false. Given the recent publication of that study, as well as the whole of the scientific literature suggesting an embodied basis of language perception, this is our least desirable conclusion, and the least likely to be assumed. The second is that, yes, we do not get an interaction effect of modality and contrast in a blocked design, and hypothesis 3 is correct. The final conclusion is that there should be an interaction effect, and that our stimuli were ill-suited to the design and prevented the observations we should have seen.

We performed some analyses of the stimulus items for clues as to the reliability of our result. If there was no feature of the stimuli that inhibit the interaction effect, then we still can't conclude that hypothesis 3 is correct pending the result of a future non-blocked study. We found some potential issues with the stimuli such that if this non-blocked study does not obtain an interaction effect, and maybe even if it does, we will motivate any revision to the stimuli in future studies according to these issues.

There was a huge skew of average accuracy per stimuli item. Some word pairs had perfect accuracy in all trials, and others were very close. There were stimuli that globally had extremely poor accuracies. The worst of these was "pale dawn." Also, valid word pairs were more likely to be incorrectly responded to, while invalids were much more commonly seen on the higher end of the accuracy by item distribution. Both of these effects likely result from the strategy that subjects were given for deciding validity. Because of the existence of metaphorical pairs (howling wind) as well as early complaints about the ambiguity of the categories we asked our subjects to assign either pair to, for the blocked study we began using an explicit set of instructions. Subjects were asked to consider the second, object, word literally, and to decide if the

first word is a property of the literally taken object. The example students received to distinguish between common pairings and valid relationships was “gummy bear:” the pair described a real object, one that you can even bring to mind, but it was emphasized that this pair would not be considered valid because literal bears are not gummy. Instructions consisted of the examples “tiny planet,” “tall dwarf,” “furry kitten,” and “angry kitten.” Tiny planet and tall dwarf, even though sensical, and in the case of the former very often evoking a unique case, would be invalid pairs, as planets are not tiny, and dwarves are not tall. Kittens are furry but, despite the possibility, in general, kittens are not angry, and so these examples were valid and invalid, respectively. This was, in general, the strategy we wanted the subjects to undertake; however, we may have omitted a relevant point, and made a mistake with one example. Some stimuli feature properties that are rarely, or occasionally true of the target noun, and were classed as valid in our study. This example with “angry kitten,” was therefore an inappropriate instruction to give. Subjects seem to have classified possible, occasional, or rare, but valid, properties of the noun as invalid, reserving valid only for the essential, intrinsic, or extremely probable properties of the corresponding objects. This strategy explains the skew towards invalid versus valid pairs, as word pairs likes “muffled screams,” valid on our design, were perceived as something like “not necessarily valid,” and so were classified invalid. In addition, the emphasis on a literal take on the target object might have generalized so that metaphorical or non-literal word pairs would be classified invalid. This explains why “pale dawn,” “hazy dust,” and “howling wind” were observed with such low overall accuracies. The fact that stimulus features seem to correspond with the instructed strategy suggests that strategy might have had an effect on reaction time as well. In future studies, we’ll make sure to include the possibility for both metaphorical language (on the construction level) and the validity of “occasionally” valid properties, on top of the strategy noted.

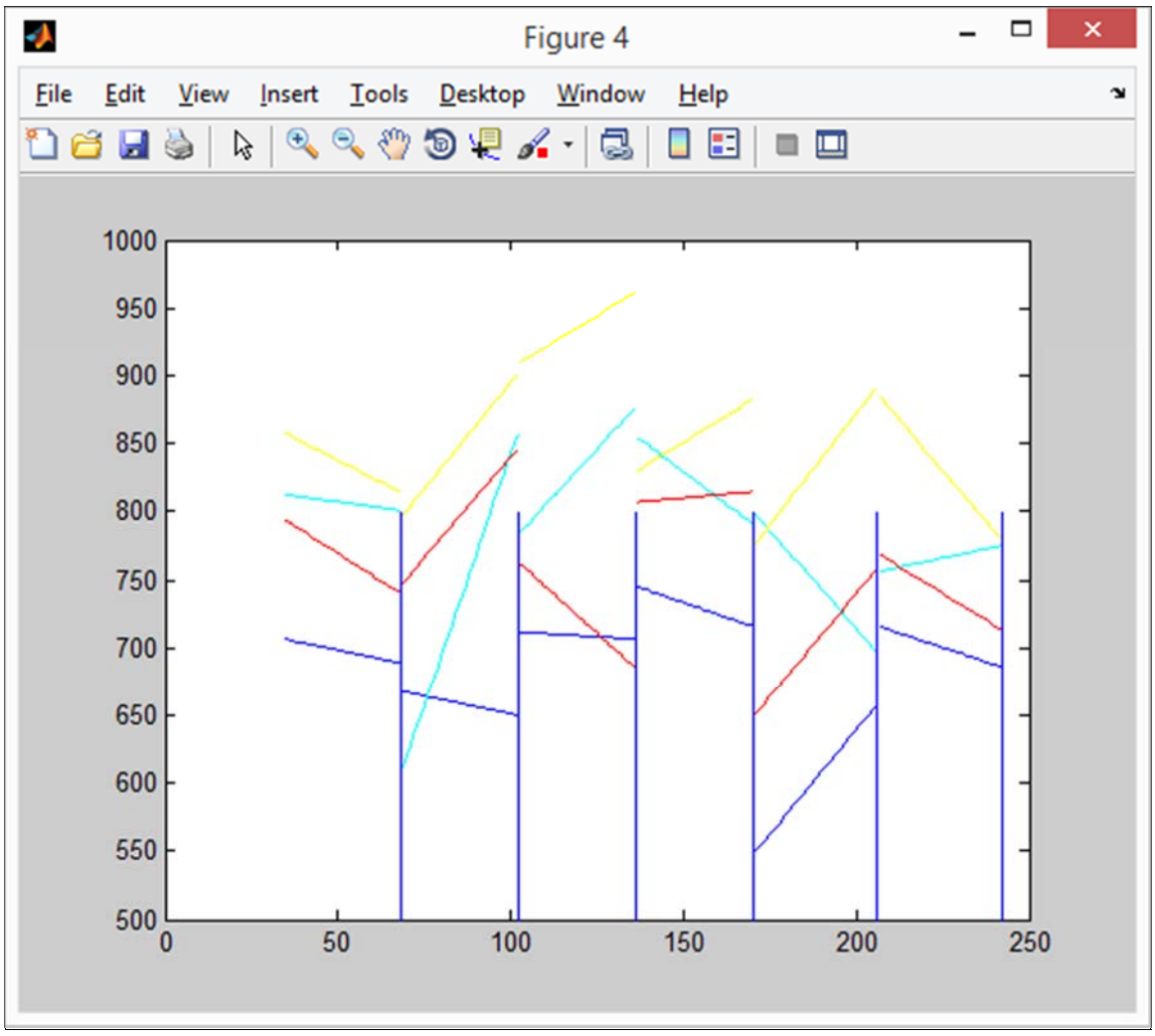
The item accuracy list can be found at: <https://docs.google.com/spreadsheets/d/1heBaEz6N9i8Nj7VjJmIlCS_wJb_JzS0wzJImBUcnp14/edit#gid=719074703>

We found that there was a main effect of modality on accuracy, though it was only a difference of about 3% accuracy. There was no interaction effect of modality and contrast on accuracy.

Another concern is that a large number of our auditory word pairs have strong visual aspects or affects

associated with them. “Screaming newborn,” listed as auditory, was, in independent questioning of both subjects and associates, associated more with an image of a newborn’s crumpled, red face, mouth agape and arms flared, than any sound of the child’s crying. “Giggling girls,” “howling hounds,” “shrieking victims,” and “roaring lion” all produce similar concerns. For future studies a subject questionnaire or a pre-study survey are all options to minimize the visual aspect of auditory stims, and I will be sure to consider them then.

We plotted reaction time by stimuli presentation order, and regressed to fit lines the figure below. This plot is color coded, such that dark colors are high contrast, light are low, cool colors are auditory, warm are visual. This plot shows that, in the first block of an experiment, on average, subjects will keep a pretty consistent reaction time, and improve slightly throughout the block. On subsequent blocks, the reaction time is significantly more erratic. Subjects varied heavily in their overall improvement within a block, suggesting some sort of effect of blocking on reaction time. Since low contrast reaction times were especially affected, this suggests the existence of an interaction effect between block order and modality.



The best explanation for this phenomena is a long term habituation effect across blocks. This entails that a subject's habituating to a certain contrast level has some long-term effect on his or her ability to perceive stimuli in subsequent blocks. This suggests that some kind of switch effect still maintained in the blocked design. Since our basic premise is to disambiguate the mechanisms or processes of the interaction effect seen in the original ERP study, this issue needs to be addressed in future studies of this paradigm. The most likely cause of this effect is the inclusion of indefinite subject-ended break screens in between any two blocks. Subjects might have been in a rush to complete their experiments, and ended these breaks very quickly. In this case, a subject that has been strongly habituated to a certain contrast suddenly switches contrast with no time to recover, producing a particularly poignant switch cost. The inclusion of any switch cost invalidates the hypotheses we can conclude based on differences between the blocked and un-blocked experiments. In light of this, for all future studies we perform with these hypotheses, the incorporation of minimal break time will be a necessary feature of any working design.

Conclusion

We ran a blocked design experiment in keeping with previous studies, intending to observe the interaction effect of contrast vs modality on reaction time of validity judgements for word pairs. Our study was meant to work in tandem with an otherwise identical study differing only in the inclusion of blocked modality. This difference, in principle, will omit any effects of switching between modalities in short time frames from the interaction observed, and the result, given the successful implementation of this single difference, will allow for the disambiguation between potential mechanisms underlying the interaction effects observed in the successful ERP study described above. This study did not obtain an interaction effect of modality and contrast on reaction time, and we could not accept any prescribed hypotheses, as all three required a successful obtaining of an interaction effect in a study identical except for the blocking by contrast design. In addition, we discovered several concerns that undermine the reliability of our own design. These are an issue with strategy skewing results, visual affects in auditory pairs confounding our results, and an indicator of some potential switching effects despite our design. I have noted resolutions to all these concerns, and will see to controlling them on future studies.