# Ling 245 Class Project Paper

Benjamin Sparkes

June 8, 2018

## 1 Introduction

The core question of Bott and Chemla (2016) is whether or not there are shared reasoning processes which apply to distinct instances of enrichment via alternatives, or whether each category of enrichment has its on specialised process. (2016, 118) One may worry that this begs the question with respect to what kind of explanation one should give to pragmatic reasoning of the kind considered by Bott and Chemla in favour of enrichment via alternatives. There are two items to note in this respect. First, assuming that an explanation is correct is not quite the same as begging the question. So long as we, or Bott and Chemla, are not attempting to support enrichment via alternatives as an explanation for (certain cases of) pragmatic reasoning, there's no worry in assuming that it's correct. Second, Bott and Chemla, indeed, are not attempting to support enrichment via alternatives as an explanation for (certain cases of) pragmatic reasoning. This may not appear to be consistent with statements such as: 'Our approach was to test whether enrichments can be primed across expressions.' (2016, 118). This may suggests that Bott and Chemla are interested in whether priming can occur at all, but when this is followed by: 'If different sorts of enrichments can prime each other, there must be an abstract mechanism that is shared between them. By testing which enrichments prime each other and which don't, we can specify what the common mechanism might be.' (2016, 118) It should be clear that Bott and Chemla are interested in whether priming can occur at all *given* a prior assumption of enrichment via alternatives.

The question of whether or not there are shared reasoning processes which apply to distinct instances of enrichment via alternatives, or whether each category of enrichment has its on specialised process has a nice cognitive feel. For, it would seem that there's some positive upshot whichever way the data points. If there is cross-category enrichment, then there is a need to posit shared reasonig processes. And, if there is no cross-category enrichment, then one should posit distinct reasoning processes. While the question has a nice cognitive feel given these resolutions, it is important to note that each of these carries a presupposition that the data can/should/will support one of these resolutions. However, (and as we shall see) there is no guarantee that the data will be so clean.

Bott and Chemla run three experiments, in which participants are presented with trials consisting of a setence and two picutures, and are asked to select the picture which best reflects the sentence. We'll go into the details below, as for now the relevant detail is that trials are split into *prime* and *response*, and every response trial is preceeded by two prime trials which are used to ensure the participant considers certain alternatives. For the two pictures in the response trial, one picture is consistent with the semantic content of the sentence, and the other contains the words 'Better Picture?', which the participants were instructed to click if they felt the the other picture did not sufficiently capture the sentence meaning. This allows us to state the basic linking hypothesis, which is that prior trials will effect how participants evaluate sentences, and that in response trials participants click on 'Better Picture?' if they process the setence pragmatically, and the semantically adequate picture otherwise.

Bott and Chemla do not make predictions regarding the results of the experiment. For, their core interest is in how the question noted above should be resolved. However, from the results of Bott and Chemla's experiment, one should expect to see a significant effect of priming, both within and between categories.

This paper contains the results of a (partial) replication of Experiment 1 of Bott and Chemla (2016). The replication is partial for two reasons: 1) the replication ran with half the number of participants compared with Bott and Chemla's original experiment (100 and 200 participants, respectively), and 2) the replication contained only two enrichment categories, as opposed to three in the original. The basis for both modifications was straightforward cost considerations, and by uncommenting a few lines of code (and fixing any bugs that this may cause) allows for the full experiment to be run. We will discuss the second aspect of this modification in detail after reviewing Bott and Chemla's paper.

The code for the experiment, data collected, analy-

sis scripts, and other relevant resources can be found at https://github.com/bsparkes/bottchemla2016[1], and one can experience the experiment at https://bsparkes.github.io/bottchemla2016/experiment/html/bottchemla2016.html.

The experiment was registered with OSF, though due to forgetfulness this was not strictly a *pre*registration as the experiment had been initialised earlier the same day. However, as the analysis of the experiment will follow that of Bott and Chemla, there isn't much room for funny business. The registration is available at the following url: https://osf.io/5bnmr/register/5771ca429ad5a1020de2872e.

# 2 The Experiment

## 2.1 Method

*Participants*

One hundred participants were recruited using Amazon Turk. Following Bott and Chemla we removed 7 participants who did not declare English as their native language, and the data from the remaining 93 participants were used in the experiment.

Further, we included keyboard shortcuts to help participants complete the experiment, where the left or right card could be selected by pressing the left or right arrow, respectively, and this could be confirmed by pressing on the space bar. This meant that in principle the participants could complete the experiment very quickly. For example, going through the experiment as fast as possible (using the keyboard, not reading the sentnces, etc.) takes around 40 seconds. One and a half second seems a reasonable lower bound for time spent on a trail[2], which would require participants to spend at least three minutes on the experiment, excluding time spent on instructions and other tasks. So, we excluded a single participant who fell below this lower bound.

In the discussion section we will explore relaxations and restrictions of this constraint.

*Materials*

Each trail involved a sentence presented above two pictures. Participants were asked to select one of the two pictures which best reflects the sentence. The sentence was constructed using one of two frames: (i) Some of the symbols are [symbol] (ii) There are four [symbol] Bott and Chemla included a third frame: (iii) There is a [symbol]. As mentioned in the introduction, this frame was excluded for cost considerations. We shall keep track of the differences to the experiment which follow from using two frames as opposed to three in this section, and engage in a broader discussion in the Discussion portion of this paper.

The symbols were one of diamonds, clubs, ticks, spades, hearts, squares, stars, circles, notes, or triangles. Pictures consisted of rectangles in the style of playing cards which contained either symbols of the test "Better Picture?". In prime trials both pictures contained symbols, while from target trials the left picture contained symbols and the other the "Better Picture?" text.

Pictures which contained symbols could be strong, weak, or false. Strong prime trails involved a strong and a weak picture. Weak prime trails involved a weak and a false picture.

For each prime trials there was a 'correct' response, either due to the semantic content of the sentence in the case of weak trials, or due to pragmatics in the case of strong trials. As Bott and Chemla write 'in the presence of both a weak picture and a strong picture, participants could not make a non-arbitrary choice solely based on the truth conditions of the weak interpretation which is true in both cases, hence the strong reading is a favored option in that it provides a non-arbitrary way to resolve the task.' (2016, 124)

In *some* trials strong pictures involved three symbols matching the predicate in the sentence, and six of another type. For example, the picture corresponding to the sentence "Some of the symbols are spades" would be three spades and six of instances of some other symbols, such as diamonds. Bott and Chemla do not specify how these symbols are arranged, and so we randomised between a line of three symbols matching the predicate at the top of the picture, and at the bottom of the picture. Weak pictures involved nine symbols matching the predicate in the sentence, and false pictures involved nine symbols of the same type which did not match the predicate.

In *number4* trials strong pictures involved symbols matching the number and predicate in the sentence, the number was always 'four'. For example, the picture corresponding to the sentence "There are four circles" would be four circles. Weak pictures involved a greater number of symbols than in the sentence which matched the predicate, following Bott and Chemla this was always six. False pictures involved a smaller number of symbols than in the sentence which matched the predicate, following Bott and Chemla this was always two.

Details for *ad hoc* trails can be found in Bott and Chemla (2016, 123–124). In addition to ad hoc trials, Bott and Chemla included *ad hoc bias* trials at the start

[1]Though https://gitlab.com/bsparkes/bottchemla2016 is more likely to stick around.

[2]After restricting by language, the mean completion time was just under 9 minutes.

Table 1: Experiment 1 results from Bott and Chemla (2016, 125).

| | | $\beta$ | S.E. | $Z$ | $p$-value |
|---|---|---|---|---|---|
| Overview | Prime $*$ WithBet + (1 + Prime $*$ WithBet \| subject) | | | | |
| | (Intercept) | -0.594 | 0.198 | -2.991 | .003 |
| | Prime | 0.563 | 0.034 | 16.342 | <.001 |
| | WithBet | 0.126 | 0.029 | 4.284 | <.001 |
| | Prime:WithBet | -0.430 | 0.033 | -13.177 | <.001 |
| Within simple | Prime | 0.993 | 0.059 | 16.950 | <.001 |
| Between Simple | Prime | 0.133 | 0.033 | 4.082 | <.001 |
| Within detail | Prime $*$ WithCat + (1 + Prime $*$ WithCat \| subject) | | | | |
| | (Intercept) | -2.088 | 0.255 | -8.185 | <.001 |
| | Prime | 1.239 | 0.109 | 11.374 | <.001 |
| | WithCatNUM4 | 2.068 | 0.195 | 10.588 | <.001 |
| | WithCatSOME | 1.823 | 0.157 | 11.598 | <.001 |
| | Prime:WithCatNUM4 | 0.174 | 0.166 | 1.046 | .269 |
| | Prime:WithCatSOME | -0.138 | 0.137 | -1.007 | .314 |
| Between detail | Prime $*$ BetCat + (1 + Prime $*$ BetCat \| subject) | | | | |
| | (Intercept) | -0.691 | 0.204 | -3.384 | <.001 |
| | Prime | 0.145 | 0.058 | 0.058 | .012 |
| | BetCatSOMEADH | -0.054 | 0.089 | -0.611 | .540 |
| | BetCatSOMENUM4 | 0.889 | 0.112 | 7.915 | <.001 |
| | Prime:BetCatSOMEADH | -0.069 | 0.079 | -0.873 | .383 |
| | Prime:BetCatSOMENUM4 | 0.078 | 0.088 | 0.888 | .374 |

*Note*. R-pseudo code shown in the first line of every section. *Prime* = priming factor (2 levels: strong, weak). *WithBet* = within/between factor (2 levels: within, between). *WithCat* = within expression category factor (3 levels: *some*, number4, *ad hoc*). *Betcat* = between expression category factor (3 levels: *some* $\leftrightarrow$ number4, *some* $\leftrightarrow$ *ad hoc*, number4 $\leftrightarrow$ *ad hoc*).

of the experiment. To quote Bott and Chemla; 'The idea behind the bias trials was to facilitate participants in imagining what the appropriate "better picture" might be for the enriched expression.' (2016, 124) As we did not include ad hoc trials we did not include these ad hoc bias trials.

*Design*

There were two types or enrichment category (*some* and *number4*), and for each category there were two prime and target types (*strong* and *weak*. So, there were $2 \times 2 \times 2 = 8$ distinct prime-target combinations, *prime* $\rightarrow$ (*strength* $\times$ *target*). Following Bott and Chemla there were four examples of each prime-target combination, so there were 4 (examples) $\times$8 (prime-target combinations) $\times$3 (triplets) = 96 experimental trials, or 32 experimental triplets.

In contrast, as Bott and Chemla included *ad hoc* trials, and do there were $3 \times 2 \times 3 = 18$ distinct prime-target combinations, and so 4 (examples) $\times$18 (prime-target combinations) $\times$3 (triplets) = 216 experimental trials, or 54 experimental triplets.

Bott and Chemla included a further 36 filler trials, 12 per enrichment category. So, there was one filler trial for every 6 target trials. To keep this ratio between filler and target trials we included 15 filler trials. This gives a filler trial for every 6.4 target trials.

*Randomisation and 'counterbalancing'*

Following Bott and Chemla all participants saw the same set of target trials, though as we included fewer filler trials than there were filler trial types, we took two filler trial types from the *many* category, two from *number*6, and an additional from either filler type *many* or *number*6 chosen at random for each participant. The symbol in the sentence and the pictures was always chosen at random for each trial. Prime-target triplets had a distinct construction as discussed above, however the order of these triplets was randomised for each participant (both target and filler triplets were included in the randomisation).

As noted above, for each prime trial there was a 'correct' response, and the position of this correct response was randomised. This contrasts with Bott and Chemla who ensured that the position of the correct response was counterbalanced across trials so that in half of the

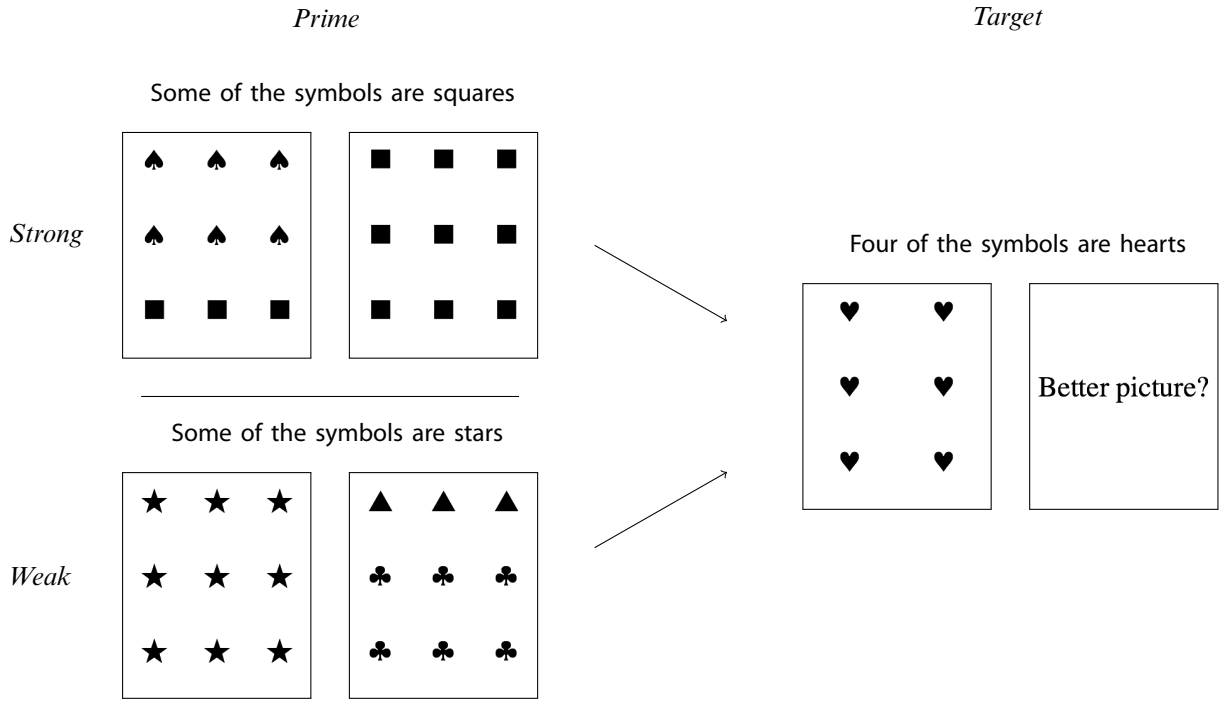


Figure 1: *Example stimuli for the replication.* Participants see two instances of a prime type followed by a target. The prime (left) consists of a sentence and two pictures, and the target (right) consists of one picture containing symbols and a 'Better Picture?' option. Here, the schema for a cross category triplet is shown, when the prime is taken from the *some* category, and the prime from the *number*4 category. The symbols used were generated randomly, and the outlines for each picture had curved corners which I was too lazy to reproduce in tikz.

trails it was to the left, and in half to the right and that in half of the trials the correct response was the same side as the previous trial and in the other half it was on the opposite side (2016, 124). So, again we have not quite exactly replicated Bott and Chemla's experiment, but Bott and Chemla only specify that the position of the correct picture was counterbalanced, and do not, for example, say that this counterbalancing was spread evenly across prime-target triplets, was held fixed across participants, etc. Rather than think through a series of design choices with unclear details and motivation, randomisation of placement on each trial for each participant seemed far more straightforward. However, in the case of target trials we followed Bott and Chemla in always placing the "Better Picture?" option to the right (2016, 124).

*Procedure*

- Note the keyboard shortcuts added, and the help screen at the start.

## 2.2 Results

> Look back at some of the readings if you're not certain what goes in each section. If you're doing a replication, include in Methods the ways in which you deviated from the original or weren't able to completely reproduce the original (e.g., because of lack of information or because you only chose to run a subset of conditions, etc.). If you're doing a replication, also include in Results the extent to which you replicated the original result(s). Include intuitive visualizations of the data in the Results section.

Each target trial was preceded by two prime trials. Bott and Chemla use this design to filter out target responses where they cannot be sure that the participant understood the correct interpretation of the prime sentence. For Bott and Chemla this led to the removal of 875 out of 13,360 target responses (2016, 124). In our replication the same procedure led to the removal of 216

[3] Unlike Bott and Chemla's results for the experiment 1, we keep the between category priming groups distinct (see Bott and Chemla (2016, Fig. 2, 122)). However, Bott and Chemla also present distinct results for *some* and *number*4 respones when compiling the results from all three of their experiments, and so the panels may be compared to these (see Bott and Chemla (2016, Fig. 6, 133)).

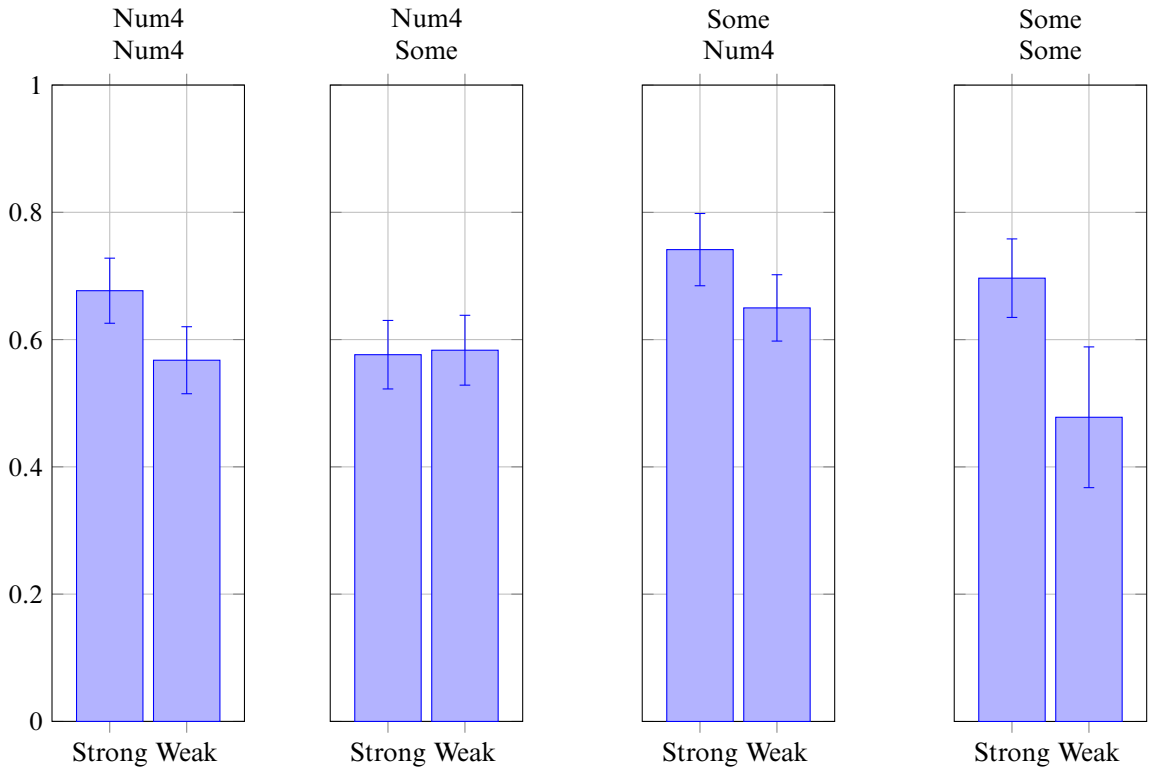


Figure 2: *Replication results*. Priming is shown by the difference between the strong and weak bars for each panel. The label at the top of each panel shows the prime and response types. For example, the third panel, labelled 'Some, Num4' corresponds to priming with *some* and a *number*4 response.[3] It is unclear what the error bars for the corresponding panels are for Bott and Chemla, here the error bars correspond to 90% confidence intervals.

out of 2976 target responses. In terms of a comparison of relative target response removals, the numbers are 6.5% and 7.2% of all trials, respectively. Bott and Chemla note that a slightly larger number of *some* trials were removed in comparison to *ad hoc* and *number* targets (2016, 124) and as we did not include *ad hoc* trials this may explain the slight difference between the experiments. However, as Bott and Chemla do not include information about the categories the incorrect primes were removed from, we don't have sufficient information to establish this explanation in any robust sense.

# 3 Discussion

Begin by briefly summing up the motivating question and main results and end on a brief concluding paragraph. In between:

If you're doing a replication: (to what extent) did the original results replicate? Discuss potential reasons for any differences, and any other qualms you may have with the design or other aspects of the experiment.

If you're not doing a replication: to what extent were the predictions borne out? If not borne out, what are some potential reasons why?

# References

Bott, Lewis and Emmanuel Chemla (2016). "Shared and distinct mechanisms in deriving linguistic enrichment". In: *Journal of Memory and Language* 91, pp. 117–140.

Table 2: Experiment 1 results from our replication.

| | | $\beta$ | S.E. | $Z$ | $p$-value |
|---|---|---|---|---|---|
| Overview | Prime $*$ WithBet + (1 + Prime $*$ WithBet \| subject) | | | | |
| | (Intercept) | 0.962 | 0.346 | 2.778 | <.01 |
| | Prime | 0.310 | 0.074 | 4.196 | <.001 |
| | WithBet | -0.006 | 0.067 | -0.089 | .929 |
| | Prime:WithBet | 0.294 | 0.071 | 4.135 | <.001 |
| Between Simple | Prime | 0.016 | 0.089 | 0.181 | .857 |
| Within Simple | Prime | 0.603 | 0.114 | 5.277 | <.001 |
| Within detail | Prime $*$ WithCat + (1 + Prime $*$ WithCat \| subject) | | | | |
| | (Intercept) | 1.361 | 0.460 | 2.960 | <.01 |
| | Prime | 0.759 | 0.206 | 3.678 | <.001 |
| | WithCatSOME | -0.784 | 0.432 | -1.816 | .069 |
| | Prime:WithCatSOME | -0.164 | 0.265 | -0.618 | .536 |
| Between detail | Prime $*$ BetCat + (1 + Prime $*$ BetCat \| subject) | | | | |
| | (Intercept) | 0.899 | 0.506 | 1.777 | .076 |
| | Prime | -0.086 | 0.160 | -0.541 | .589 |
| | BetCatSOMENUM4 | 0.861 | 0.451 | 1.910 | .056 |
| | Prime:BetCatSOMENUM | 0.362 | 0.282 | 1.281 | .200 |

*Note*. R-pseudo code shown in the first line of every section. *Prime* = priming factor (2 levels: strong, weak). *WithBet* = within/between factor (2 levels: within, between). *WithCat* = within expression category factor (3 levels: *some*, number4, *ad hoc*). *Betcat* = between expression category factor (3 levels: *some* $\leftrightarrow$ number4, *some* $\leftrightarrow$ *ad hoc*, number4 $\leftrightarrow$ *ad hoc*).

| Overview | | $\beta$ | S.E. | $Z$ | $p$-value |
|---|---|---|---|---|---|
| | Prime + (1 + Prime \| Subject) | | | | |
| Some $\rightarrow$ number4 | Prime -0.080 | 0.162 | -0.492 | 0.623 | |
| number4 $\rightarrow$ some | Prime | 0.354 | 0.238 | 1.488 | 0.137 |

Table 3: etc.

# A Additional Data

| Prime Type | Prime Category | Response Category | Percentage mean | Raw mean | Raw S.D. | Raw S.E. |
|---|---|---|---|---|---|---|
| Strong | Num4 | Num4 | 0.6767956 | 2.634409 | 1.653619 | 0.1714723 |
| Weak | Num4 | Num4 | 0.5675553 | 2.184783 | 1.683334 | 0.1745536 |
| Strong | Num4 | Some | 0.5762712 | 2.193548 | 1.702032 | 0.1764925 |
| Weak | Num4 | Some | 0.5833029 | 2.239130 | 1.750162 | 0.1814834 |
| Strong | Some | Num4 | 0.7414502 | 2.511364 | 1.597371 | 0.1656396 |
| Weak | Some | Num4 | 0.6498584 | 2.466667 | 1.643510 | 0.1704240 |
| Strong | Some | Some | 0.6966165 | 2.329545 | 1.713514 | 0.1776831 |
| Weak | Some | Some | 0.4703510 | 1.978261 | 1.728737 | 0.1792617 |

Table 4: Details for the plots contained in Figure 2.