Ling 245 Class Project Paper

Benjamin Sparkes

June 15, 2018

Abstract

Partial replication of Experiment 1 from Bott and Chemla (2016). Main results are: a) Replication of priming effects in general, b) failure to replicate between-category priming effect, c) replication of within-category priming effect, Main theoretical result: less support for shared reasoning processes between distinct instances of enrichment via alternatives.

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 (EVAs), 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.

In any case, 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 reasoning 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 ran three experiments, in which participants are presented with trials consisting of a setence and two pictures, 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 preceded 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. Instead, their core interest is in how the question noted above should be resolved. However, they do note that '[i]f enrichment can be primed at all, we would expect within-category priming' and that '[i]f the *numbers*, *some* and *ad hoc* EVAs share enrichment mechanisms we would expect

them to prime each other, so that a strong some prime, for example, leads to a greater proportion of strong number responses.' (2016, 122) And, 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, analysis 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 (`https://osf.io/5bnmr/register/5771ca429ad5a1020de2872e`). Though, due to forgetfulness this was not strictly a *pre*registration as the experiment had been initialised earlier the same day. Still, as the analysis of the experiment will follow that of Bott and Chemla, there isn't much room for funny business.

2 The Experiment

2.1 Method

Materials

The experiment consistent of trails, where each involved two pictures presented below a sentence. 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 later in this paper.

The symbols were one of diamonds, clubs, ticks, spades, hearts, squares, stars, circles, notes, or triangles.[3] Pictures consisted of rectangles in the style of playing cards which contained either symbols or the text "Better Picture?". In prime trials both pictures contained symbols, while in target trials the left picture contained symbols and the other "Better Picture?".[4]

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

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

[2] 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 responses 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)).

[3] The following Unicode symbols were used: ♦ ♣ ✓ ♠ ♥ ■ ⋆ ● ♩ ▲. One participant noted initial confusion in terming '✓' as 'tick' instead of 'checkmark', in further experiments one may wish to use a different symbol as the others do seem very clear.

[4] In Bott and Chemla's example stimuli the "Better Picture?" option contained a darker background, but this was not mentioned in the text, and we could see no clear motivation for doing so. Instead, the option had the same background as all other pictures—white.

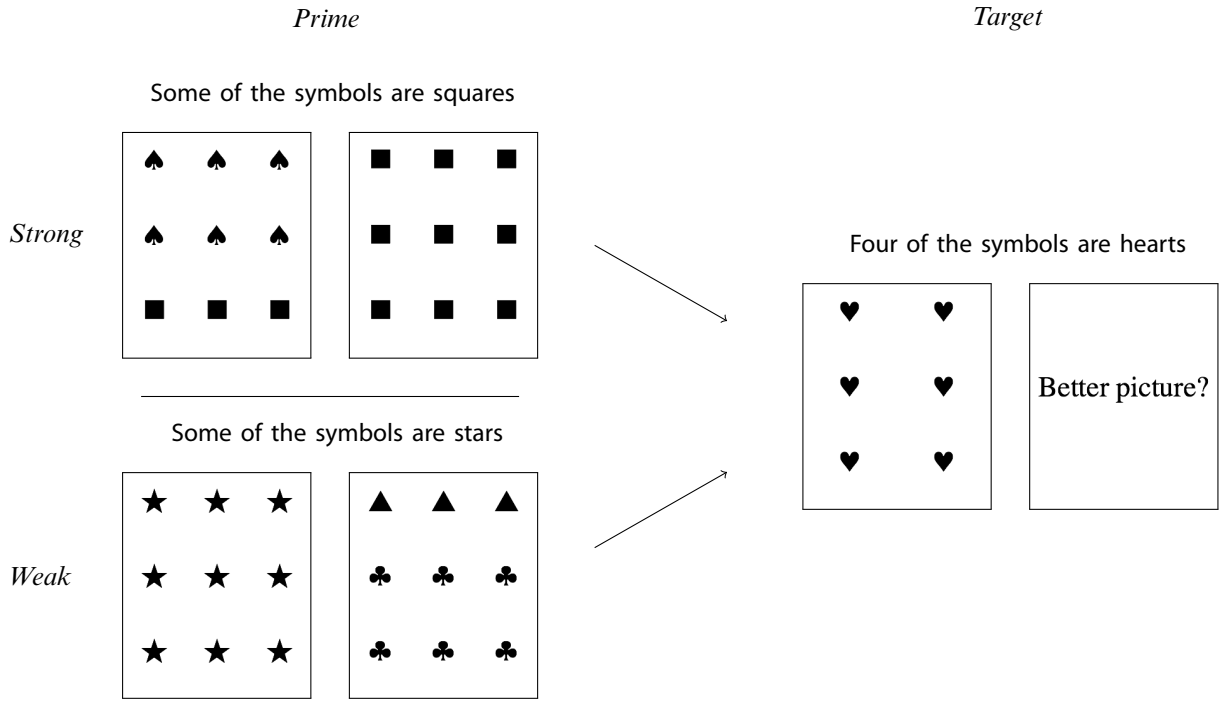

Figure 1: *Example stimuli for the replication.* Participants see two instances of a prime type followed by a target. The prime (left) consists of a sentence and two pictures, and the target (right) consists of one picture containing symbols and a 'Better Picture?' option. Here, the schema for a cross 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.

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 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.

Filler trials were also included. There were *all* sentences, an alternative to *some*, and *number*6 sentences, an alternative to *number*4. For example, "All the symbols are [symbol]" and "Six of the symbols are [symbol]". Each could occur in three forms: (1) a weak picture with symbols that did not match the predicate in the sentence, and a "Better Picture?" option (2) a weak picture with symbols that matched the predicate, and a "Better Picture?" option, and (3) a weak picture with symbols that matched the predicate, and a strong picture. Bott and Chemla used these to highlight alternatives to participants.

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 so 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. Bott and Chemla note that filler trials were 'linked to each prime-target combination' (2016, 123), but did not specify whether this link was simply conceptual, or also part of the ex-

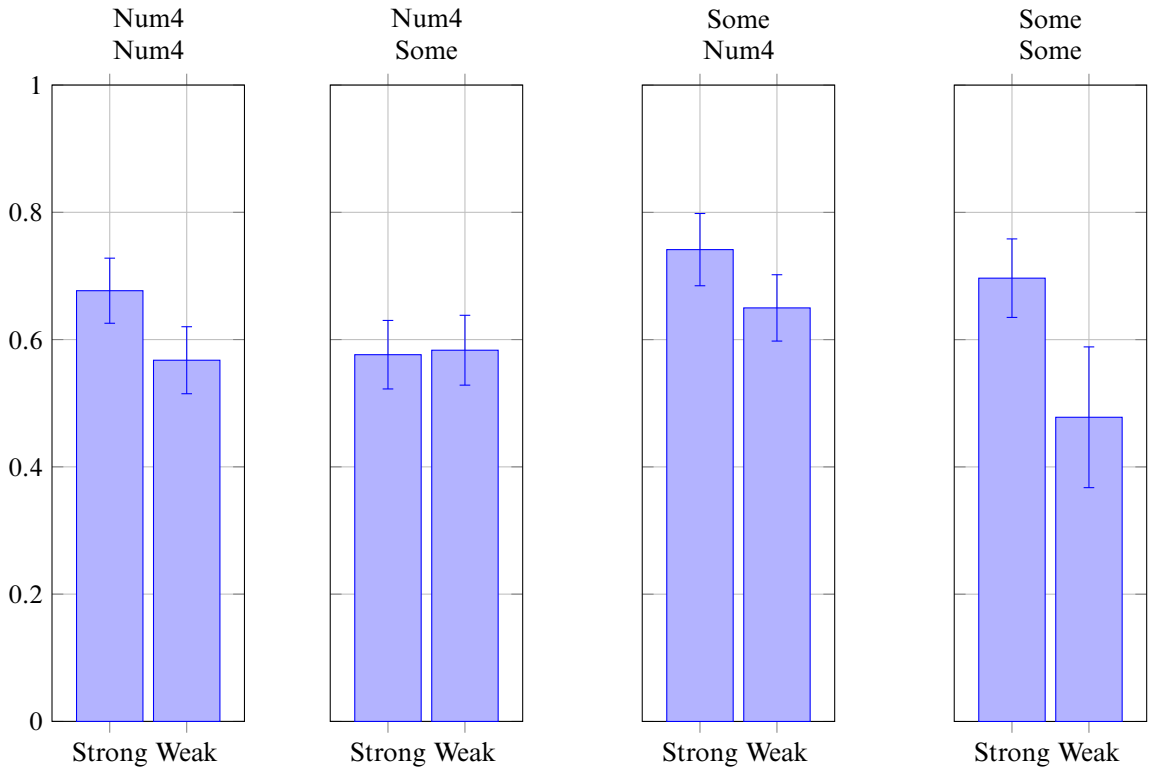

Figure 2: *Replication results*. Priming is shown by the difference between the strong and weak bars for each panel. The label at the top of each panel shows the prime and response types. For example, the third panel, labelled 'Some, Num4' corresponds to priming with *some* and a *number*4 response.[2] 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.

periment. It is for this reason that we distributed the filler trials randomly.

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 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 on the right (2016, 124).

Procedure

As noted, Participants were instructed to click on the picture that "best reflected the sentence", and were given one example with the sentence "Many of the symbols are [symbol]" and another with the sentence "There is a [symbol 1] above a [symbol 2]". The latter included

a picture for which sentence was false and the "Better Picture?" option, with a reminder as to when it may be appropriate to click the "Better Picture?" option.[5]

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 functionality was detailed at the start of the experiment, and participants were able to test what pressing a key would do). 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 sentences, etc.) takes around 40 seconds. One and a half second seems a reasonable lower bound for time spent on a trail[6], 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.[7]

2.2 Results

Data treatment

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 194 out of 2,750 target responses. In terms of a comparison of relative target response removals, the numbers are 6.5% and 7.1% 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.

Analysis procedure

Follow Bott and Chemla the response-type likelihood was modelled using logit mixed-effect models. Analyses were conducted using lme4 (Bates et al. 2014), languageR (Baayen 2011), and memisc (Elff 2012), libraries for the R statistics program (Team et al. 2013). The data is presented in the same format as Bott and Chemla with β values, standard errors, Z-values, and p-values shown in the tables accompanying the experiment together with R pseudo-code describing the models. Treatment and sum coding were used as described by Bott and Chemla, with any factor not explicitly mentioned receiving treatment contrasts. The memisc package was used to ensure that both types of contrasts had the same bases as in the original experiment. The random effects structure included random intercepts and slopes for all repeated measure factors.

The analysis starts with an general model involving all of the data, in which the interaction of with and between-expression priming is assessed. A more detailed analysis is then performed by restricting the analysis to within and between-expression trials only. The dependent measure was the log odds of choosing a strong over a weak prime.

Analysis

Fig. 2 shows the results from the replication, and the corresponding figure can be found in Bott and Chemla (2016, 122). Table 1 reports statistical details of the original experiment, and Table 2 reports the same from the replication. These figures and tables are fairly self explanatory, but a detailed walk-through can be found in Bott and Chemla (2016, 125).

[5]It is unclear whether or not this differs from Bott and Chemla, as they do not note the contrast to the "Better Picture?" option. And, one may argue that using an example where the sentence would be straightforwardly false on the other option (the relevant symbols were next to each other) to illustrate the use of "Better Picture?" may prejudice the participants in favour of a semantic interpretation. In the former example one of the cards contained 9 symbols, with all but one matching the sentence symbol, and the other contained 9 symbols where only 3 matched the sentence symbol. And again, this could reasonably be taken to require a semantic interpretation of the sentence. Still, as the interest of the experiment is whether pragmatic enrichment can be primed, an initial semantic bias should not be too much of a concern. Participants were asked to 'think again' if they selected the false picture in both of the examples.

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

[7]This is perhaps an odd mix of trial-by-trail exclusion, and broad participant exclusion. It would have been interesting to depart from Bott and Chemla's approach and exclude participants who failed to correctly respond to some certain percentage of primes (say 15% or so). Perhaps some other time.

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

		β	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 [base]). *WithBet* = within/between factor (2 levels: within [base], between). *WithCat* = within expression category factor (3 levels: *some*, number4, *ad hoc* [base]). *Betcat* = combined between expression category factor (3 levels: *some* ↔ number4, *some* ↔ *ad hoc*, number4 ↔ *ad hoc* [base]).

Bott and Chemla report three analyses: 1. whether EVAs can be primed at all, 2. whether priming occurs at the within-category level, and 3. whether priming occurs at the between-category level. While these analyses are combined in the above tables, appendix B contains separate tables for each analysis which combine the results from the original experiment and the replication.

For the model used in the first analysis, sum contrasts are used for both factors, and a significant effect of a strong prime increasing the rate of strong responses, $\beta = 0.56$, $p < .001$, a significant effect of strong responses happening in between category trials, rather than between category, $\beta = 0.126$, $p < .001$ and an interaction between the two $\beta = -0.43$, $p < .001$ showing that the effect of the prime was greater in between category trials.

We observed slightly different results. First, a significant effect of priming $\beta = 0.310$, $p < .001$, no significant effect of between category trials, $p = .929$, but a significant effect of the interaction between the two $\beta = .294$, $p < .001$ showing that the effect of prime was greater for within category trails. Though, in all cases these effects were smaller than observed by Bott and Chemla. A direct comparison can be seen in Table 6.

Bott and Chemla use a model with a similar structure, but using treatment contrasts for the within/between factor and sum contrasts for the prime factor to investigate simple effects. Bott and Chemla observed significant priming occurred at the within category level, $\beta = .99$, $p < .001$, and at the between category level $\beta = .13$, $p < .001$. We observed significant priming at the within category level $\beta = .603$, $p < .001$, but no significant priming at the between category level $\beta = .016$, $p < .857$.

So, while Bott and Chemla observed priming of EVAs at the within-category level and the between category level, we only observed a significant effect of priming at the within-category level.

Bott and Chemla broke the data down into within-category trials and between-category trials to assess the observed effects in more detail, conducting separate analyses on each. These constitute the remaining two analyses. In each model treatment contrasts were used for the categories, and sum contrasts for the prime.

For within category trials no significant difference between *some* and *number*4 categories was observed $\beta = 2,07$, $p < .001$ and $\beta = 1.74$, $p < .001$, respectively, with *ad hoc* as the base category. In contrast, *number*4

Table 2: Experiment 1 results from the replication.

		β	S.E.	Z	p-value
Overview	Prime $*$ WithBet + (1 + Prime $*$ WithBet | subject)				
	(Intercept)	0.962	0.346	2.778	<.010
	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	<.010
	Prime	0.759	0.206	3.678	<.001
	WithCat	−0.784	0.432	−1.816	.069
	Prime:WithCat	−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
	BetCat	0.861	0.451	1.910	.056
	Prime:BetCat	0.362	0.282	1.281	.200

Prime = priming factor (2 levels: strong, weak). *WithCat* = within category factor (2 levels: *some*, *number*4 [base]). *Betcat* = between category factor (2 levels: *some* → number4 [base], *some* → *number*4).

served as a base in our model, and *some* differed from this $\beta = -0.78$, $p = .069$. Not quite a significant effect, but a suggestive difference. A direct comparison can be seen in Table 7.

Between categories, Bott and Chemla found a significant effect only for *some*/*number*4 trials, but not when interaction with the strength of the prime was accounted for, and we found no significant interaction.

However, in the case of between category trials, no direct comparison of results should be made. This is because Bott and Chemla pool between category trials (e.g., a *some* prime and *number*4 target would be treated the same as a *number*4 prime and *some* target) while we do not. This reason for this is that our replication contained only two categories, and so between category trails could not be meaningfully pooled.

Still, a meaningful comparison can be made with further results that Bott and Chemla report. For, Bott and Chemla perform additional analyses of the data pooled from three experiments they conducted (2016, 132–133), looking at the separate prime and target categories in the case of between category trials (see Table 4). Bott and Chemla observe a similar and significant effect of priming in the cases of a *some* prime and *number*4 target and a *number*4 prime and *some* target, with βs of .345 and .221, respectively. We did not observe a significant effect, even if one wished to ignore this, we did not observe a similar effect, with βs of −.080 and .354, respectively. A direct comparison of these results can be seen in Table 9.

In general the between category effects Bott and Chemla observed have not been replicated.

Bott and Chemla also examine splits of the data into first and second halves of the experiment. In line with Bott and Chemla the split by halves did not reveal a difference in responses (see Bott and Chemla (2016, Table 4, 134) and Table 5 of this document).

3 Discussion

The motivation for the experiments Bott and Chemla performed was to see what kind of priming processes there were, and whether these were distinct or shared. The original results suggest some kind of shared priming processes for the categories we considered in our replication. And, while our results are not directly comparable (see above for an explanation), this does not appear to have been replicated.

Now, Bott and Chemla go into some further detailed reasoning about the βs, and ps of the models they obtained in order to support their conclusion (see 2016, 126,132–134). In short, similar slopes are taken to indicate similar underlying processes. We won't repeat this reasoning here for the simple reason that we are sceptical as to whether such detailed reasoning *should* be considered, given the difference in results between Bott and Chemla's original experiment and our own. For, as we noted at the start, Bott and Chemla's question relied on the presupposition that there are either shared or distinct mechanisms for pragmatic processing.

Yet, the failure of replication suggests either that there were differences in the experimental setup which were not accounted for, that there were relevant background details which the experimental setup did not control for, or that there are uniform pragmatic processes which are shared between speakers in general.[8]

With regards to differences in experimental setup, it may be important that we did not include *ad hoc* trials, and in particular a number of *ad hoc* bias trails at the start of the experiment (intended to improve the overall response to *ad hoc* trials in Bott and Chemla's original experiment, see (2016, 124)). However, if the difference in observations is tied to this, then it would seem we are assuming some shared process in pragmatic reasoning between *some*, *number*4, and *ad hoc* trials, while Bott and Chemla take the results of their experiment to show that there are shared mechanisms between *some* and *number*4, but not *ad hoc*, primes (see 2016, 125–126).

It is also the case that Bott and Chemla appear to have used a grey background for their "Better Picture?" picutre (see 2016, Fig. 3, 123), while we used a white picutre (see Fig. 1). However, it is not clear why this should make a significant difference, and Bott and Chemla do not given any explanation.

Further, we did include keyboard shortcuts. Perhaps requiring participants to click on a button with their mouse affected how much attention they paid to each picture, or affected the experiment in some other way. In hindsight, it would have been useful to separate trials completed using mouse clicks from those completed using keyboard shortcuts. It is not clear whether the above observation could have be supported by this separation, but this data would have been relatively easy to obtain, and may have yielded some insights.

The number of participants between experiments also differed, though given that we observed significant effects for within category priming, it seems doubtful that this should affect the overall results. Though, another possibility is that the pay for participants differed, and assuming some strong correlation between amount of pay and the attention of participants this may be relevant, though this seems quite the assumption. Bott and Chemla do not report what they paid participants, while we paid participants $.50. This may have been a little too low, considering the task, but on a small trial most participants responded that this was a fair price (and as noted, this experiment was conducted under somewhat tight finances). Still, in both the trial and experiment a number of participants did report that they felt the pay was too low.

Along similar lines, Bott and Chemla were not clear on how the symbols in the pictures were to be laid out, and different layouts may have been easier to process. We haven't detailed in full how the symbols were laid out either, but have noted above certain relevant concerns, and the interested reader can read through the javascript code for generating images, or run an number of trial experiments to observe the differences as they would appear to participants.

The placement and number of filler trials could also be a consideration. Bott and Chemla used filler trials to highlight alternatives to participants (2016, 123). While the ratio of filler trials was approximately the same, the absolute number was not. Further, if Bott and Chemla were correct in filler trials highlighting alternatives, placement of filler trials may have been important. Unfortunately, Bott and Chemla do not state how they were dispersed through the experiment, and so we chose to place them randomly for each participant.

Finally, the length of the experiment may have been a factor. Bott and Chemla's experiment was roughly twice the length of the replication, and it may be argued that this was important for some reason. Perhaps as participants became more familiar with the task their responses shifted. Or, participants gave more immediate responses in the longer experiments, and more thought-through response in the shorter trial. It's unclear whether participants were aware of the length of the task in the original experiment, but in our retrial participants could see a progress bar. However, following Bott and Chemla we did perform an analysis of the first and second halves of the experiment, split per individuals, and in line with Bott and Chemla found no significant difference between either half, so it seems doubtful that this is an important factor to consider.

4 Conclusion

The aim of Bott and Chemla (2016) was to better understand how people use alternatives to enrich the basic meaning of a sentence. Bott and Chemla observed significant effects of priming both within and between categories, and in particular between *some* and *number*4 categories. In our replication we found significant effects of priming within categories, but no significant effects of priming between categories, even when the prime and target categories were controlled for. What this means for the issue of whether there are shared or distinct pragmatic processes is unclear, and we have tentatively explored whether the results are due to differences in experimental setup, and whether the assumption that there are either shared or distinct pragmatic processes should be explored further.

[8] In other words, either there were relevant details of the experimental setup which we did not reproduce, there were relevant details which should have been considered as part of the experiment which we did reproduce, or the presupposition that there are either shared or distinct processes requires more careful consideration.

References

Baayen, R Harald (2011). "languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"". In: *R package version* 1.1.

Bates, Douglas et al. (2014). "Fitting linear mixed-effects models using lme4". In: *arXiv preprint arXiv:1406.5823*.

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

Elff, Martin (2012). "memisc: Tools for management of survey data, graphics, programming, statistics, and simulation". In: *R package version 0.95-38, URL http://CRAN. R-project. org/package= memisc*.

Team, R Core et al. (2013). "R: A language and environment for statistical computing". In:

A Additional Data

Prime		Response	From the replication				From Bott and Chelma	
Type	Category	Category	mean %	Raw mean	Raw S.D.	Raw S.E.	mean %	Raw S.E.
Strong	Num4	Num4	0.6767956	2.634409	1.653619	0.1714723	0.615	0.018
Weak	Num4	Num4	0.5675553	2.184783	1.683334	0.1745536	0.339	0.018
Strong	Num4	Some	0.5762712	2.193548	1.702032	0.1764925	0.553	0.019
Weak	Num4	Some	0.5833029	2.239130	1.750162	0.1814834	0.484	0.019
Strong	Some	Num4	0.7414502	2.511364	1.597371	0.1656396	0.544	0.020
Weak	Some	Num4	0.6498584	2.466667	1.643510	0.1704240	0.474	0.019
Strong	Some	Some	0.6966165	2.329545	1.713514	0.1776831	0.604	0.019
Weak	Some	Some	0.4703510	1.978261	1.728737	0.1792617	0.340	0.018

Table 3: Details for the plots contained in Figure 2. Relevant cell mean and S.E. from Bott and Chemla included (see Table A1 (2016, 138–139)).

		β	S.E.	Z	p-value
	Prime + (1 + Prime | Subject)				
some → *number*4	Prime	−0.080	0.162	−0.492	0.623
*number*4 → *some*	Prime	0.354	0.238	1.488	0.137

Table 4: Analysis of priming effect for between category trials.

		β	S.E.	Z	p-value
Within by half	Prime ∗ WithCat ∗ Half + (1 + Prime ∗ WithCat ∗ Half | Subject)				
	(Intercept)	−0.221	0.378	−0.584	.559
	Prime	0.141	0.320	0.442	.658
	WithCat	1.068	0.481	2.222	<.050
	Half	0.850	0.371	2.294	<.050
	Prime:WithCat	0.770	0.542	1.423	.155
	Prime:Half	0.172	0.243	0.706	.480
	WithCat:Half	−0.671	0.334	−2.011	<.050
	Prime:Withcat:Half	−0.534	0.379	−1.407	.159
Half 1 only	(Intercept)	0.675	0.3157	2.139	<.050
	Prime	0.314	0.1154	2.719	<.010
	WithCat	0.377	0.1801	2.092	<.050
	Prime:WithCat	0.217	0.1975	1.098	.272
Half 2 only	(Intercept)	1.518	0.6091	2.493	<.050
	Prime	0.519	0.2572	2.016	<.050
	WithCat	−0.203	0.299	−0.680	.497
	Prime:WithCat	−0.392	0.361	−1.086	.277

Table 5: Analysis of the experiment by halves. Half = experiment half factor (2 levels: first half, second half). First half and *number*4 as bases.

B Direct Comparisons

		β	S.E.	Z	p-value
Overview	Prime $*$ WithBet $+$ (1 $+$ Prime $*$ WithBet | subject)				
Original					
	(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
Replication					
	(Intercept)	0.962	0.346	2.778	<.010
	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
Within Simple	Prime	0.603	0.114	5.277	<.001
Between Simple	Prime	0.016	0.089	0.181	.857

Table 6: Table contained results of original experiment and replication for the overview models.

		β	S.E.	Z	p-value
Within detail	Prime $*$ WithCat $+$ (1 $+$ Prime $*$ WithCat | subject)				
Original					
	(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
Replication					
	(Intercept)	1.361	0.460	2.960	<.010
	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

Table 7: Table contained results of original experiment and replication for the within models.

		β	S.E.	Z	p-value
Between detail	Prime $*$ BetCat + (1 + Prime $*$ BetCat | subject)				
Original					
	(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
Replication					
	(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:BetCatSOMENUM4	0.362	0.282	1.281	.200

Table 8: Table contained results of original experiment and replication for the between models. Note, these aren't really comparable.

		β	S.E.	Z	p-value
	Prime + (1 + Prime | Subject)				
Original					
some → *number*4	Prime	0.345	0.069	5.140	<.001
*number*4 → *some*	Prime	0.221	0.065	3.412	<.001
Replication					
some → *number*4	Prime	−0.080	0.162	−0.492	0.623
*number*4 → *some*	Prime	0.354	0.238	1.488	0.137

Table 9: Table contained results of original experiment and replication for the directional models.