Predicting collective benefit from the language of interaction partners

Malte Lau Petersen / au540041, Cognitive Science, Århus University

Exam paper for Social and Cultural Dynamics

Abstract

How do dyads that successfully collaborate coordinate their communication? Some tasks, such as coming up with words or foraging, can easily be parallelized, and the results from many individuals can simply be added together. Other tasks require collaborative effort in an ongoing social interaction to benefit from multiple people working on the same problem. But why do some groups benefit more from collaboration than others? Language is central for collaboration, and various mechanisms have been proposed to explain differences in collective benefit. This paper extends a previously reported experimental setup (Bahrami et al. 2010, Fusaroli et al 2012a, Bang et al. 2014, Fusaroli & Tylén 2015, Bang et al 2017) to the auditory domain – and compares linguistic similarity and diversity, linguistic alignment and synergy, and explicit metacognition as predictors of collective benefit in a bayesian computational model.

Keywords: *social cognition; group performance; linguistic alignment; bayesian modelling; auditory discrimination*

Introduction

Our ability to work together towards a common goal is an important human skill. One of the big questions is why we do it at all. We pay opportunity costs, take risks and sacrifice our own potential gains for the benefit of others. Sometimes we can hold a reasonable expectation that we will be paid back down the line – like an insurance policy or karma system. But even when there's no reason to think our actions will have any consequences – for instance in an anonymous economic game such as the trust game – we tend to have a pro-social bias. That is weird from an economic-rational perspective, put it seems to have paid off. This tendency is sometimes called altruism, and it can confer and advantage to our genes from an evolutionary point of view. We share half our genes with our siblings, and I think an argument could also be made that a group of organisms can benefit from a culture of collaboration

But a lot of this field of inquiry leans on the assumption that collaboration is costly, and that we should try to find out why we're willing to pay the price. On the other hand, we have a common intuition that *two heads are better than one*. Or we say that *the whole is greater than the sum of its parts*.

That leads me to the question: *When* does collaboration pay off? The answer depends on at least three kinds of factors: The task at hand, our relative skill levels, and the way we collaborate in order to reach the goal. This paper is mostly about the third factor, but first I will briefly touch on the other two.

Humans thrive in very diverse environments and solve many different kinds of problems both on a planet scale and in our daily lives so the fact that some problems lend themselves better to direct collaboration than others should come as no surprise. If we are foraging for a rare plant or ringing doorbells for a cause, having two people walk over the same ground might not be nearly as efficient as having each person cover a different area. But if we are trying to move bulky office tables, having one person lifting in each end can simplify the problem many times over.

Another thing that can make collaboration inefficient is large skill differences. In 2010, Bahrami et al published a series of experiments exploring the relationship between individual performance differences and collective benefit showing that in a visual perceptual decision-making task, dyads were able to out-perform the best individual in the dyad only when they were similar in skill. When the reliability of one of the dyad members was drastically reduced by introducing noise to their stimuli, the dyads' joint decisions became worse than the best individual. They argue that this deterioration is due to neither social loafing, interpersonal competition or groupthink, but that it's instead a consequence of the strategy we use for combining perceptual information across individuals. Bang et at (2017) recently showed that dyad partners tend to align their confidence ratings. They argue that this equality-bias explains why dyads with large skill differences perform worse – the better performing dyad member is willing to listen to the less reliable member even when they know they've been inaccurate in the past.

Language and collaboration

There are probably other important factors in group work, but it's pretty clear that communication is key. In Bahrami et al's 2010 paper, dyads were able to reach a collective benefit even without any explicit feedback about their performance – but only when they were allowed to communicate.
But which parts of language are relevant for differences in collective benefit, and can we measure these linguistic behaviours with automated language processing?
The following features of the language use of interaction partners were extracted and all expected to contribute to predicting the collective benefit from working together.

Similarity and Diversity

We might first look at factors at the level of the individuals. If these effects turn out to be important, group work could be made more efficient by assigning people to work with the right kinds of collaborators. Similarity in this context means being of similar skill level. If a chess novice tried to collaborate with a grandmaster, neither of them would probably become any better at chess, simply because their understandings would be too far away from each other. More similar dyads would be expected to gain more from collaborating. For the visual perceptual tasks used in Bahrami 2010, this effect is well described for difference in task performance, but the same effect might be found for language use. On the other hand, we would also expect a certain kind of diversity to have a positive effect on collective benefit, and there are multiple apparent paradoxes here. The first is the seeming contradiction between similarity and diversity. But this is just due to unfortunate imprecision in the terms. If two grandmaster chess players were to collaborate (high similarity), but coming from different chess traditions, the diversity argument predicts they would gain more from collaborating than two coming from the same school of thought.

And secondly, it doesn't seem diversity can be the main driver of collective benefit because of the other paradox: Coming from different backgrounds might lead to better collaboration, but the better the collaboration, the more they might collaborate, and the more the same people collaborate, the less diverse they will become over time.

Linguistic alignment

Recently, more and more language researchers are starting to look at not just individuals, comprehending and producing language, but rather at the interpersonal dynamics that arise whenever two or more people try to communicate with each other. One of the observations is that in general, people tend to align their vocabularies to each other – If I just called my car a “ride”, theres a much higher chance that anyone who heard it would use the same term, even if they normally think of it as a “wagon” or something else. A very influential explanation was put forward by Pickering and Garrod in 2004, where they describe how automatic low-level priming could lead to alignment on multiple levels, including their situation models (individuals’ idea of *what’s going on*). In this view, aligned situation models are a prerequisite for collaboration, and indiscriminate linguistic alignment should thus predict collective benefit positively in this experiment. But such a measure might be too crude – there are many situations where perfectly aligned language would indicate less, not more aligned situation models. Fusaroli, Rączaszek-Leonardi, & Tylén (2014), note that if a student simply repeats a professor’s utterance, word for word, this probably means very little overlap in their understanding. We should expect indiscriminate linguistic alignment to be negatively correlated with collective benefit, they say, and instead look to task-specific selective alignment. In perceptual decision-making tasks such as this, we should look for alignment of the interaction partners’ confidence expressions: if the dyads can be described as trying to integrate information across multiple brains, they should be expected to share their uncertainty and try to reach a common scale for expressing this uncertainty (Bahrami et al, 2012). This can be measured both on a local scale (If interlocutor A expresses his/her uncertainty with numbers, how likely is it that interlocutor B will also use numbers in the next utterance?) and on a conversation-level scale (Out of all the confidence expressions one dyad used, how many were the most common type?)

Alignment vs Synergy

I also wanted to replicate the analysis from Fusaroli & Tylén 2015, where they used Recurrence Quantification Analysis to compare the predictive power of the existing theory of linguistic alignment against their own synergy theory of dialog. Recurrence Quantification Analysis (RQA) is a technique to quantify the recurrent structure of a time series. The question it answers can be stated as: How often does the time series repeat itself, and what are the properties of these repetitions. Not just in terms of repetitions of states (saying the same word more than once); RQA builds a multidimensional space of all the possible states of the time series, measures how often *trajectories* in this space are recurrent. An example would be the dialog: “I love you”, “I love you too” - Of all the times the first person reached “I love y..”, most of them would continue along the same trajectory with “I love you too”.

Using this measure, the linguistic alignment account would predict that when one interlocutor follows a certain trajectory (utters a specific phrase), the other would be more likely to follow a similar trajectory in the future. But, the authors claim, to fully understand conversational dynamics, we shouldn’t restrict ourselves to look at only recurrent structures where one participant repeats another. Interlocutors can also play complementary roles, such as when one asks a question that the other answers, so we should be looking at conversations as *joint actions*. The synergy model assumes that collaborators who work well together develop conversational routines, and that these routines will show up as recurrent patterns in RQA when we first merge the data from the two participants into one joint time series.

Finally, in order to show that these effects don’t just arise from each participant being more or less self consistent, they compare also to a baseline by finding the recurrent structure in each participant separately. For all of these measures, we would expect them to be positively correlated with collective benefit: The more structured the language use, the more the dyad should benefit from the collaboration.

Sensory modalities

Previous versions of this experiment (Bahrami et al. 2010, Fusaroli et al 2012a, Bang et al. 2014, Fusaroli & Tylén 2015, Bang et al 2017) have all used a visual perceptual decision-making task where participants were shown six gabor patches twice for a short time (85 ms), and the task was then to decide which of the two intervals had an oddball (slightly higher contrast). When dyads are then asked to reach a joint decision on what they saw, the prevailing interpretation is that the two interaction partners try to integrate the information from both people similar to how one person might integrate information from multiple sensory modalities, only with imperfect information transfer. If this explanation holds, it should be possible to extend the paradigm beyond the visual domain.

I think one of the reasons it might not be easy to replicate in other sensory modalities is that we are not as used to putting words to non-visual perceptual experience. Some of this can be experience-based, but most of us would probably having a hard time arguing the nuances of olfactory stimulus.

On the other hand, this might not matter if all we need to communicate is our best guess plus our confidence in that guess (Bahrami et al, 2010).

Method

In the experiment, participants sat in a classroom at a computer each with headphones on (identical computers and headphones) and were paired up into dyads with a random other person in the room so they could only communicate through the experiment. Participants knew each other (undergrad classmates), but didn't know who specifically they were paired with. They then read the instructions, were able to familiarize themselves with the stimuli and adjust the sound volume, and gave oral consent to data collection for research purposes. In each trial, they both heard the same 450 ms random vowel (e, i, o, u) sound at a random formant intensity (5, 8, 11, 14, 17, 20, 23, or 26 dB). They were then individually asked which vowel they heard, and if they answered the same, they moved on to the next trial. If they disagreed, they were put into a chat-room and asked to discuss and reach a consensus decision. In even-numbered rounds, one player had to input the decision, and in odd-numbered rounds, the other player did it. After every trial, participants were given feedback on their own, their dyad partner's, and their consensus performance. 18 participants (9 dyads) played as many rounds (mean=79.2 rounds, sd=36.5) as they could in 45 minutes. The experiment was designed to resemble the protocol presented in Bahrami et al. 2010: Experiment 1 as closely as possible.

The vowel sounds were generated with a parametric speech synthesizer (Anikin, 2016) using a base sound common to all the vowels and adding formants (overtone patterns that are the reason sounds at the same base frequency sound different) at varying intensities. At intensity = 0 dB, the four sounds would be completely physically indistinguishable, and the intensity levels were chosen in order to capture the whole range of the psychometric function. The experiment was run using oTree (Chen, Schonger & Wickens), the data was processed with R (R Core Team, 2016) and modelled using Stan (Stan Development Team, 2016).

Psychometric curves

For quantifying performance in perceptual tasks, I use a psychometric curve model (Kingdom & Prins, 2010). Separately for individuals' and dyads' data, I used a hierarchical generalized linear model with a logit link, a binomial outcome distribution, and crossed random effects corresponding to the model specification:

```
k | n ~ 1 + intensity + (1 | group) + (0+intensity | group) + (1 |
vowel) + (0+intensity | vowel)
```

Where k is the number of correctly perceived vowels out of n trials. I use a hierarchical model in order to get an estimate for the overall main effect while downplaying the impact of the structured variance accounted for by the different groups and the different vowels.

The parameter for stimulus intensity is the change in log odds of detecting the right vowel per dB of formant intensity. In other words, it represents the sharpness of the category boundary for detecting the vowel correctly. So the steeper the curve is (the higher the parameter value), the better you are are at distinguishing between the stimuli.

Following Fusaroli et al (2012), the collective benefit of a group is then defined as the ratio between the parameter value (slope) for the collective decisions to the slope for the individual decisions of the best individual in that group.

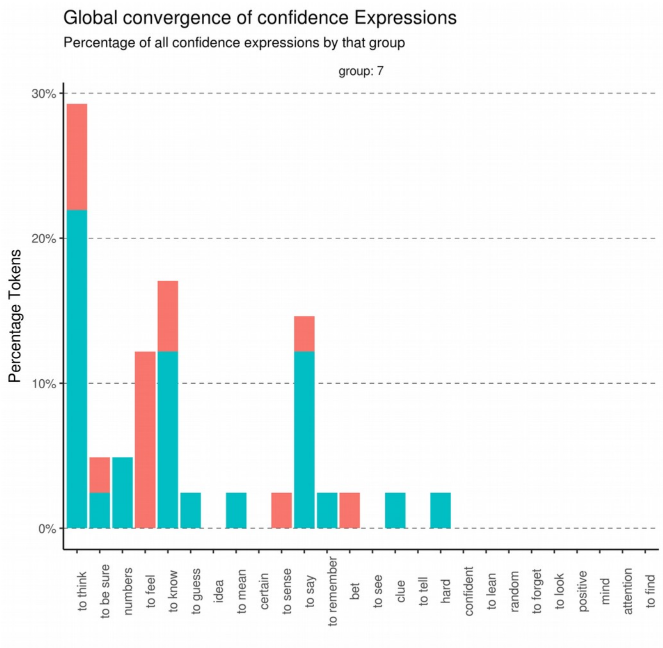
```
collective benefit = dyad slope / individual slope
```

Dialog quantification

Dyads chatted in the experiment using oTree chat, and the following text processing was done at the word level. The hypothesized benefit of similarity comes from being able to use each other well if you're at a similar level, so it was defined as the difference in vocabulary size (in the experiment) between the two dyad partners. Diversity on the other hand is supposed to help if the dyad can "cover more ground" together than alone, so it was defined as the cosine similarity between the vocabularies of the the dyad partners. The diversity argument would predict that the lower the cosine similarity between the vocabularies, the more different backgrounds the participants come from, and thus the more they will gain from working together.

Linguistic alignment was defined as the proportion of words that had been used by the other participant in the previous utterance. For indiscriminate alignment, this was done for all words, while for confidence alignment, this was run only on the subset of words that matched a category in the confidence scheme from Fusaroli et al 2012. In other words, if one participant produced an expression of confidence, what was the probability that the other participant had used a version of the same expression of confidence last time they used one?

Global convergence was defined, again following Fusaroli et al 2012, as the proportion of all confidence-expressions for each dyad that was of the most popular kind.



Distribution of global confidence-expression types of one group. Each colour is one participant. This group converged somewhat (29.3%) on "to think" as an expression of confidence. The rest of the groups are in the appendix.

Recurrence Quantification Analysis

First, words were split up into characters, which were then converted to numbers. Then I ran RQA using the crqa package in R (Moreno, Coco & Dale, 2015) with the parameters: radius = 0, delay = 1, embedding dimension = 2. This means that a recurrence is defined as a two-letter pair anywhere in the sequence that is the same as any two-letter pair in the same or another position in the sequence, and that the smallest recurrent structure that can create a diagonal line on the recurrence plot is 3 letters.

Using RQA on different parts of the chat data, I quantify three different kinds of recurrence. The baseline is individual self-consistency, which is RQA applied to each participant's corpus individually. Second, I can apply Cross-Recurrence Quantification Analysis to each group by taking each participant's corpus as an individual time-series. This quantifies the structural organization that we would expect from theories of alignment. Thirdly, if we combine both dyad partners' chat logs into one (while respecting the timings) and run RQA on that, we can quantify the structural organization of the combined cognitive system predicted by the synergy approach.

It is possible to extract a number of different measures from a recurrence plot, but following Fusaroli & Tylén 2015, I focus only on proportion of recurrent points that are part of an ongoing trajectory longer than two letters (L), and entropy of the distribution of these trajectory lengths (Entropy).

Bayesian model

To fit the full model, I use a bayesian model for several reasons: Most importantly, to my knowledge, it's not possible to do this whole analysis in a non-bayesian framework without splitting up into different parts and thus losing the ability for each part to influence the others. This leads to my second reason: building a bayesian computational model in stan allows us to carry the uncertainty all the way through the analysis to a model with the full distributions rather than the point estimates we typically use in a frequentist framework. And finally, when we do bayesian modeling, we specify our prior beliefs in the model, which guides us to being much more explicit about the assumptions our model holds. (McElreath, 2016, Chapter 6)

The model consists of three parts: The two logistic regressions for individuals and dyads, and a linear regression predicting each dyad's collective benefit from the measures we extracted from the chat data. In the plate notation, this is local confidence agreement, but the structure is the same for all the models: only the independent variables (and corresponding slope parameters) in the collective benefit part of the model changes between models. All models ran for at least 10000 warm-up iterations and 10000 samples.
Part of bayesian modeling is specifying priors for the parameters of the model. A principled approach is using skeptical priors which reduces the risk of overfitting. This corresponds to a prior belief that we can do a lot of different experiments and test for a lot of different effects, but most of them are likely to be zero or close to zero. With standardized (rescaled to mean=0, sd=1) data, a common skeptical prior for coefficients in a regression model is `normal(0, 1)`, which I've used for all parameters in the model except two: The intercept in the collective benefit model has a mean of 1 because it's a ratio and that's what corresponds to "no difference". And the variance parameter sigma in the collective benefit model is a half-cauchy with its peak on 0 and a scale of 2. The cauchy distribution has a very long tail to allow for a wide range of possible parameter values, but does still not take as much data to move as a uniform prior. Together, the priors chosen for this model should help bias the estimates in a conservative direction.

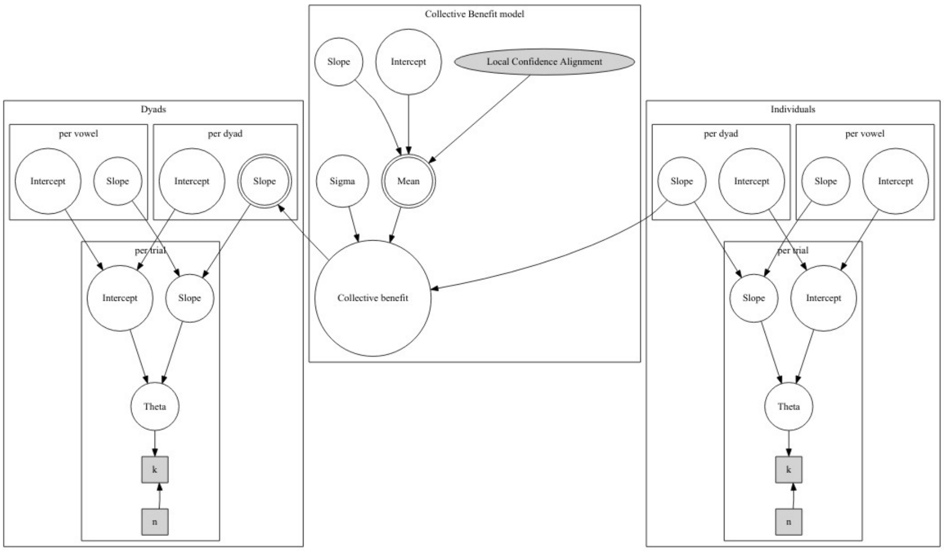


Plate notation for the model. Notice that the per dyad slope in the dyad section of the model is deterministic since it's calculated from the collective benefit in order to make the model generative. It's equivalent to collective_benefit ~ dyad slope / individual slope. This was done mostly for computational purposes[1], but it also reflects the overall hypothesis that the direction of causation is from communication to better group performance. Stan code for this model is in the appendix, and on can be found on github for the rest of the models.

Alternatively, it's possible to use the results from previous literature to inform empirical priors, but that would make hard to compare the posteriors since I didn't have access to similar quality empirical priors for all the parameters I wanted to compare.

In total I ran 12 different models on collective benefit: a null model with just an intercept, and 11 times with one extra parameter for the different measures extracted from the chat data.

Ideally, in order to minimize the risk of overfitting, we should be using out of sample error or information criteria to compare models to each other and the null model. I calculated the Widely Applicable Information Criterium (WAIC) for each model, which is the pointwise log likelihood minus a penalty based on the complexity of the model, and can be interpreted so that models with a lower information criteria is closer to the truth (McElreath, 2016). But in this case the uncertainty of these estimates are so large that it doesn't really tell us much.

1 *Discussion on the Stan mailing list: http://discourse.mc-stan.org/t/hierachical-hierachical-model-possibly-a-transformed-parameters-jacobian-issue/*

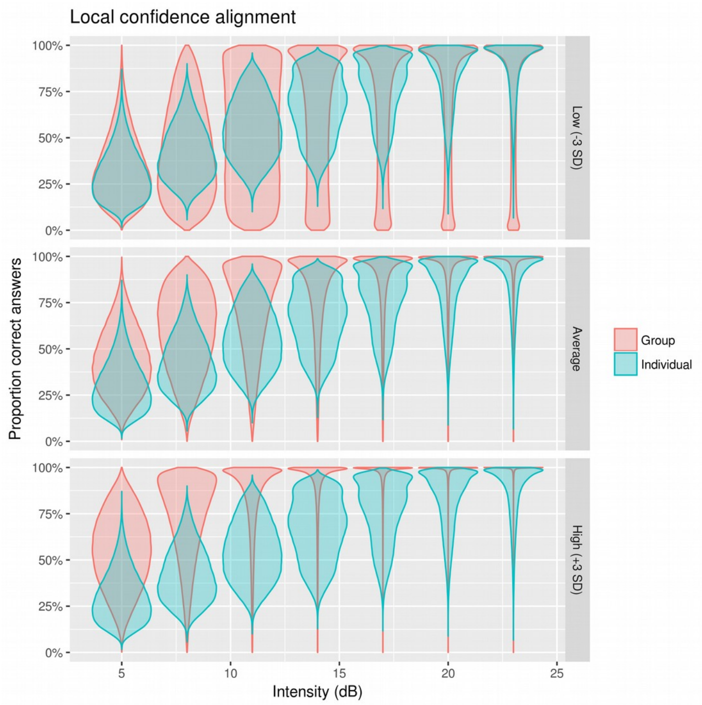
Results

I want to stress that not enough data was collected for a model this complex to know how well these results would generalize to new data.

With that said, we can plot so-called counter-factual plots of the predictive posteriors, and get the models' best predictions for the data we do have. All the hypothesized predictors except cosine similarity and local indiscriminate alignment had posteriors with a center above zero. All of these were in their expected direction.

Many of the proposed linguistic features are highly correlated, especially the RQA measures (plot of correlations in appendix). This leads to an interpretational problem, even if we were to accepts the estimates we got from these models. Due to these high correlations in the data and the fact that there was no experimental manipulation in this experiment, we can not say anything about causation. Neither the data nor the model tells us whether alignment leads to collective benefit, or if successful collaboration leads to more alignment, or if the cause is instead in something entirely different.

As to why the linguistic features inspired by different theories are highly correlated in my sample, I can see two different explanations, that we again cannot distinguish from this data. Explanation 1: They simply measure the same or very related concepts. In this case, it might be possible in the future to combine the theories that say this concept is important. Explanation 2: The same cognitive pattern gives rise to different (but correlated) linguistic behaviour. This might make it hard to distinguish between competing theories, and might even disprove (or make irrelevant) parts of them in the end.



Predictive posterior plot for the local confidence alignment model. Predictions are plotted for groups with low (mean – 3 sd), average, and high (mean + 3 sd) local confidence alignment. Note that even though the model predicts a benefit in group performance from higher local confidence alignment, the interval for the parameter crosses zero (see table below) and WAIC is inconclusive. Predictive posterior plots for the other models are in the appendix.

Model	WAIC	SE	pWAIC	dWAIC	Weight
Self-consistency entropy	1425.495	91.753	88.245	0.000	0.177
Synergy L	1425.701	91.438	87.696	0.206	0.159
Self-consistency L	1426.416	91.720	88.387	0.921	0.112
Alignment entropy	1426.661	91.802	88.508	1.166	0.099
Synergy entropy	1426.680	91.891	88.724	1.185	0.098
Local confidence alignment	1427.071	91.861	88.696	1.576	0.080
Difference in vocabulary size	1427.102	91.836	88.670	1.608	0.079
Global confidence convergence	1427.910	91.756	89.136	2.415	0.053
Null model (intercept only)	1428.185	91.575	88.434	2.690	0.046
Alignment L	1428.412	91.902	89.485	2.917	0.041
Local indiscriminate alignment	1429.054	91.892	89.964	3.560	0.030
Cosine between word sets	1429.295	91.908	89.720	3.800	0.026

Information criteria (WAIC) for all models. Lower WAIC is better. pWAIC is the complexity penalty, dWAIC is the distance to the lowest WAIC, and Weight is the relative Akaike wieght assigned to each model. The Standard errors of all the models overlap.

Model	Estimate	Lower 95%	Upper 95%
Self-consistency entropy	0.0204	-0.0057	0.0461
Synergy entropy	0.0187	-0.0121	0.0434
Self-consistency L	0.0185	-0.0081	0.0459
Synergy L	0.0159	-0.012	0.0444
Alignment entropy	0.0149	-0.0131	0.044
Local confidence alignment	0.014	-0.0108	0.0431
Difference in vocabulary size	0.0132	-0.0157	0.0423
Alignment L	0.0131	-0.0176	0.0397
Global confidence convergence	0.0127	-0.0161	0.0422
Cosine between word sets	-0.0018	-0.0331	0.0264
Local indiscriminate alignment	-0.0059	-0.0368	0.023

Maximum a posterioris and 95% highest posterior density intervals. High correlations among proposed predictors (see appendix). Models are identical except for the one slope parameter on collective benefit. All predictors were standardized to mean=0 sd=1 before entering in the model. Estimates can be interpreted as: "For every 1 standard deviation in this predictor, the dyad adds the estimate to their collective benefit".

Conclusion

I built this model in hope of being able to compare the predictive power of features calculated with different techniques, and inspired by different theoretical approaches to what drives collective benefit. I think this is the right approach to comparing theoretical approaches: Quantify their predictions and compare their predictive power with computational models. This becomes all the more important when studying the complex thing that is human interactions, and especially when the different theoretical approaches make highly similar predictions, it is not enough to test only our favorite measure.

Ultimately, this paper failed to conclusively answer the question it set out to due to a problem that's all too common in social sciences: too little data for the ambitions of the model.

Further work

The direction of this has important applications because human collaboration is so common. Science could one day provide helpful guidelines to getting the most out of working together as a group. That said, most of the research in linguistic predictors of collective benefit has been correlational in nature – this paradigm included. It might just as well be that some other factor (in the individuals or group dynamics) is the underlying cause which is driving both the patterns in language use and collective benefit. Future studies should seek to introduce an experimental manipulation to determine the causal structure. But one straightforward improvement would be to collect more data, and from different sensory modalities, different cultures, and different tasks and combine them in one model. Probably only some kinds of tasks will lend themselves to these kinds of performance improvements, and we won't know until we collect the data and run the models. Showing that the model is capable of recovering the true parameters from simulated data would also be an important step.

Open science

All data and code for the experiment, text processing, modeling, and visualizing is available on http://github.com/maltelau/SocKult-AuditoryConfidence

References

Anikin, Andrey. (2016). Parametric speech synthesizer. Downloaded from https://cogsci.shinyapps.io/shinyApp/ April 2017, but no longer available as of June 2017.

Bahrami, B., Olsen, K., Latham, P. E., Roepstorff, A., Rees, G., & Frith, C. D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-1085.

Bang, D., Fusaroli, R., Tylén, K., Olsen, K., Latham, P. E., Lau, J. Y., ... & Bahrami, B. (2014). Does interaction matter? Testing whether a confidence heuristic can replace interaction in collective decision-making. *Consciousness and cognition*, *26*, 13-23.

Bang, D., Aitchison, L., Moran, R., Castanon, S. H., Rafiee, B., Mahmoodi, A., ... & Summerfield, C. (2017). Confidence matching in group decision-making. *Nature Human Behaviour*, *1*, 0117.

Daniel L. Chen, Martin Schonger and Chris Wickens: "oTree - An Open-Source Platform for Laboratory, Online and Field Experiments", working paper available at: www.otree.org/oTree.pdf.
Chris Wickens: Otree chat plugin. Downloaded from https://github.com/oTree-org/otreechat April 2017

Pickering, M. J., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and brain sciences*, *27*(02), 169-190.

Fusaroli, R., Bahrami, B., Olsen, K., Roepstorff, A., Rees, G., Frith, C., & Tylén, K. (2012). Coming to terms: quantifying the benefits of linguistic coordination. *Psychological science*, *23*(8), 931-939.

Fusaroli, R., Rączaszek-Leonardi, J., & Tylén, K. (2014). Dialog as interpersonal synergy. *New Ideas in Psychology*, *32*, 147-157.

Fusaroli, R., & Tylén, K. (2016). Investigating conversational dynamics: Interactive alignment, Interpersonal synergy, and collective task performance. *Cognitive science*, *40*(1), 145-171.

Kingdom, F. A. A., & Prins, N. (2010). Psychophysics: A practical introduction.

McElreath, R. (2016). *Statistical rethinking: A Bayesian course with examples in R and Stan* (Vol. 122). CRC Press.

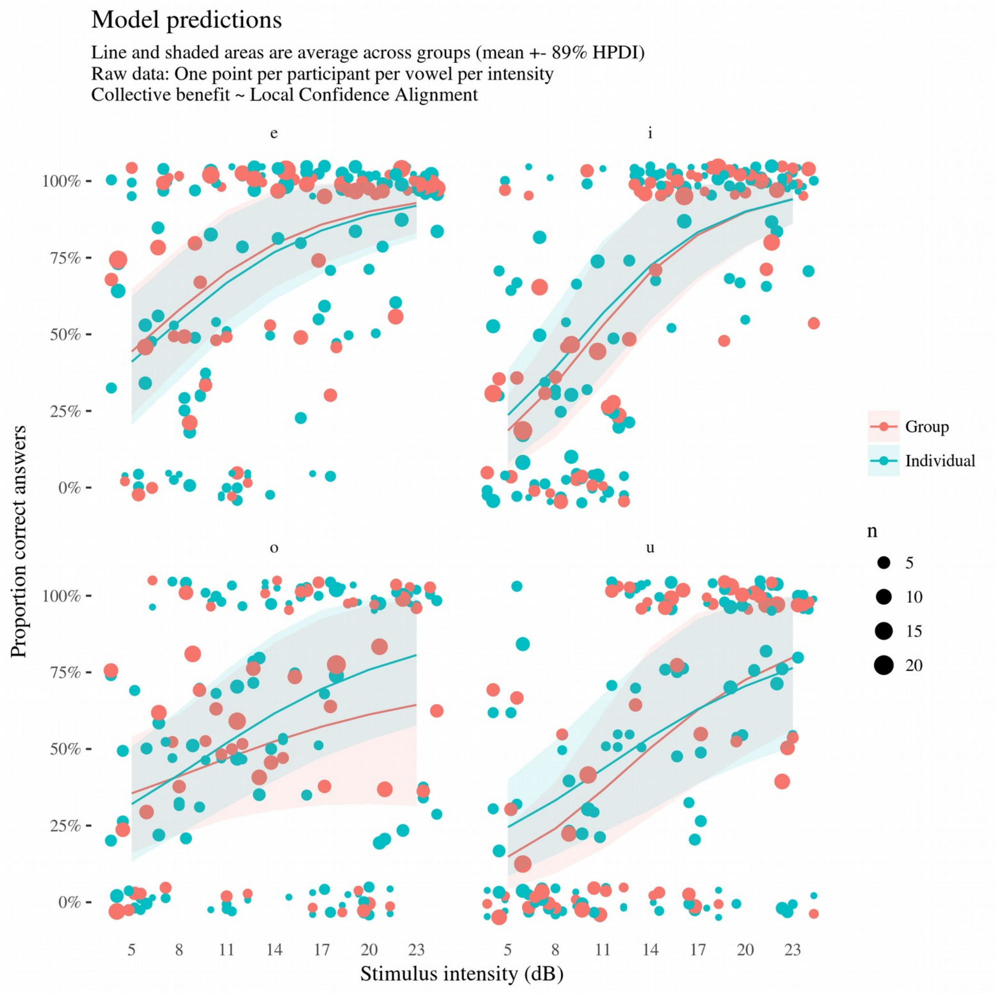
Moreno I. Coco, Rick Dale with contributions of James D. Dixon and John Nash (2015). crqa: Cross-Recurrence Quantification Analysis for Categorical and Continuous Time-Series. R package version 1.0.6. https://CRAN.R-project.org/package=crqa

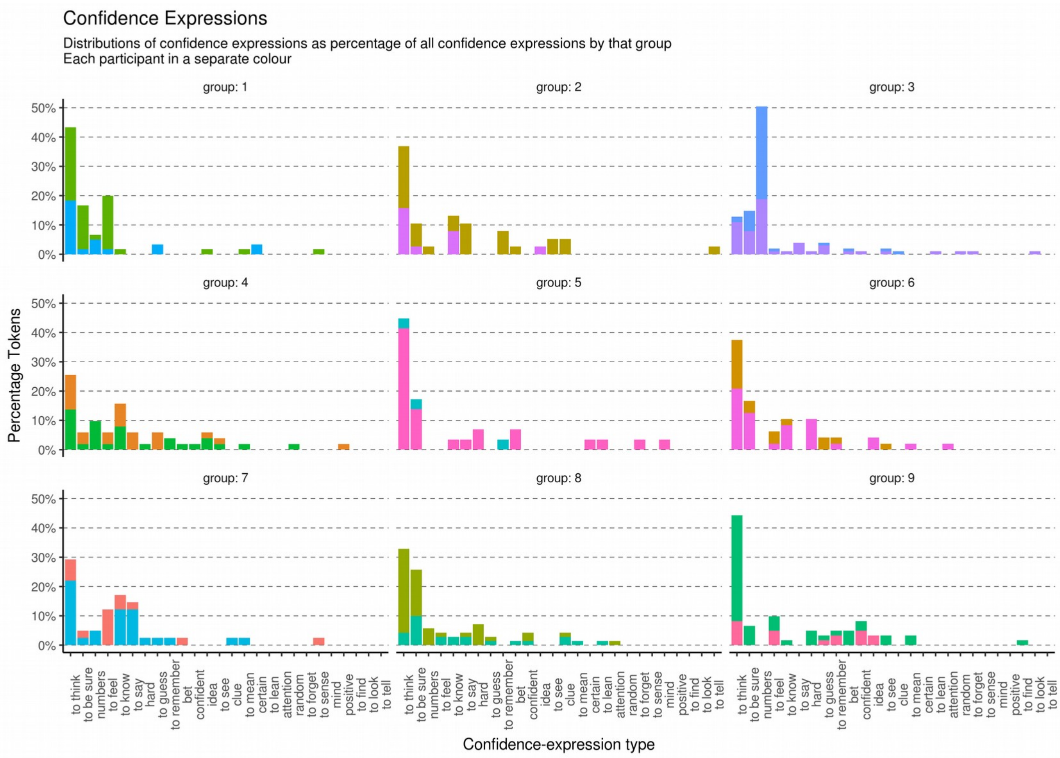
R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

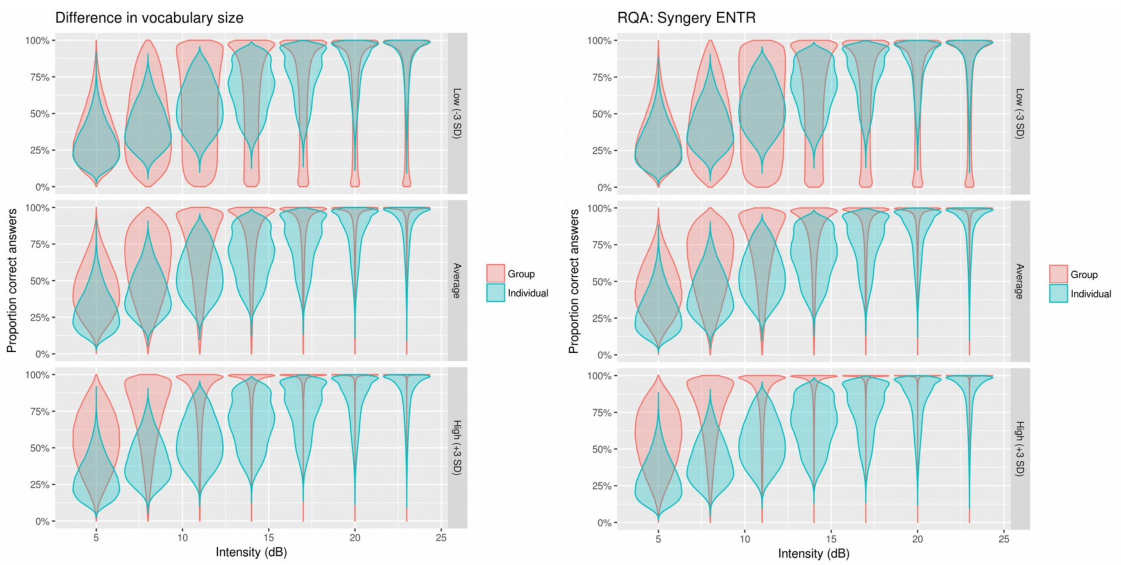
Stan Development Team. 2016. *Stan Modeling Language Users Guide and Reference Manual*, Version 2.15.0. http://mc-stan.org

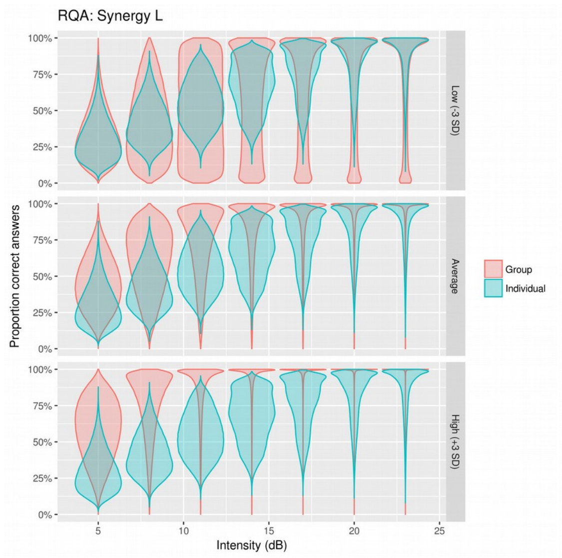
Stan Development Team. 2016. *RStan: the R interface to Stan*. R package version 2.14.1. http://mc-stan.org

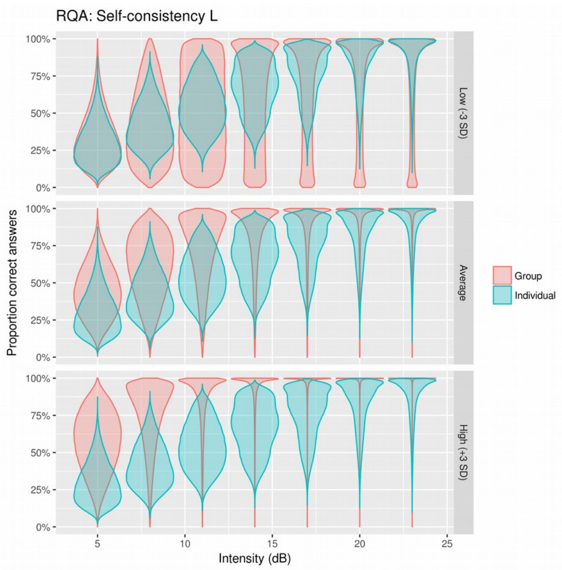
Appendix

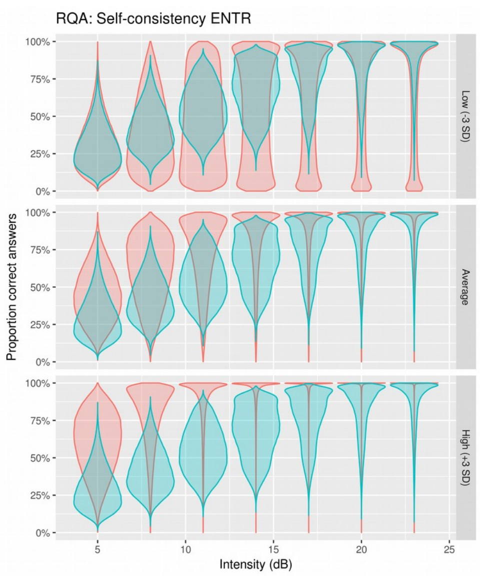

Model predictions
Line and shaded areas are average across groups (mean +- 89% HPDI)
Raw data: One point per participant per vowel per intensity
Collective benefit ~ Local Confidence Alignment
e
i
o
u
100%
75%
50%
25%
0%
Proportion correct answers
5
8
11
14
17
20
23
Stimulus intensity (dB)
Group
Individual
n
5
10
15
20

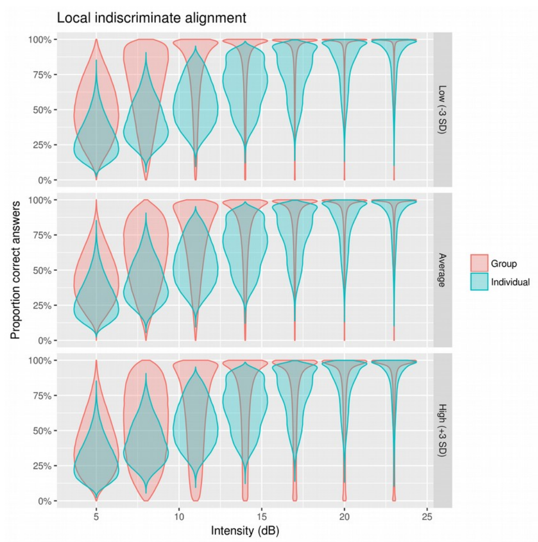

Confidence Expressions
Distributions of confidence expressions as percentage of all confidence expressions by that group
Each participant in a separate colour
group: 1
group: 2
group: 3
group: 4
group: 5
group: 6
group: 7
group: 8
group: 9
Percentage Tokens
0%
10%
20%
30%
40%
50%
to think
to be sure
numbers
to feel
to know
to say
hard
to guess
to remember
bet
confident
idea
to see
clue
to mean
certain
to lean
attention
random
to forget
to sense
mind
positive
to find
to look
to tell
Confidence-expression type

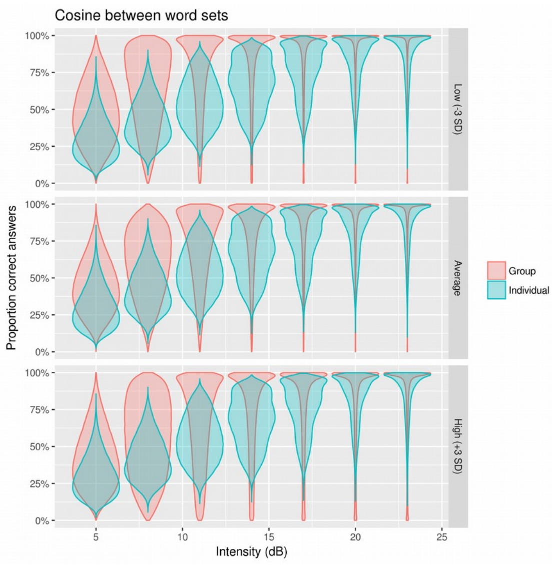

Difference in vocabulary size
RQA: Syngery ENTR
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
Group
Individual
5
10
15
20
25
Intensity (dB)

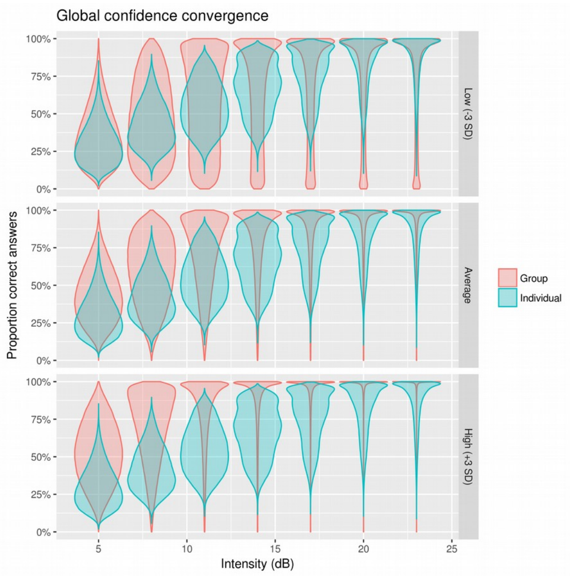

RQA: Synergy L
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
5
10
15
20
25
Intensity (dB)
Group
Individual

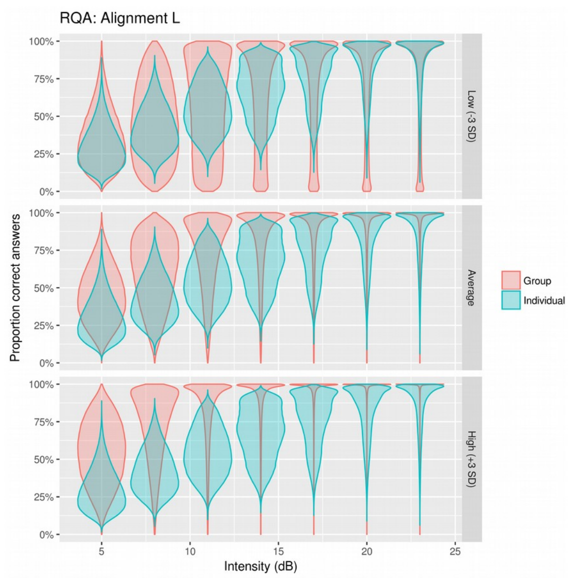

RQA: Self-consistency L
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
5
10
15
20
25
Intensity (dB)
Group
Individual

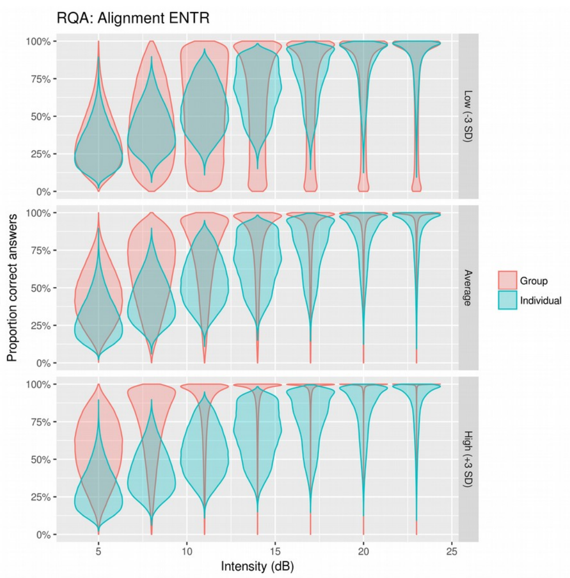

RQA: Self-consistency ENTR
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
5
10
15
20
25
Intensity (dB)
Group
Individual

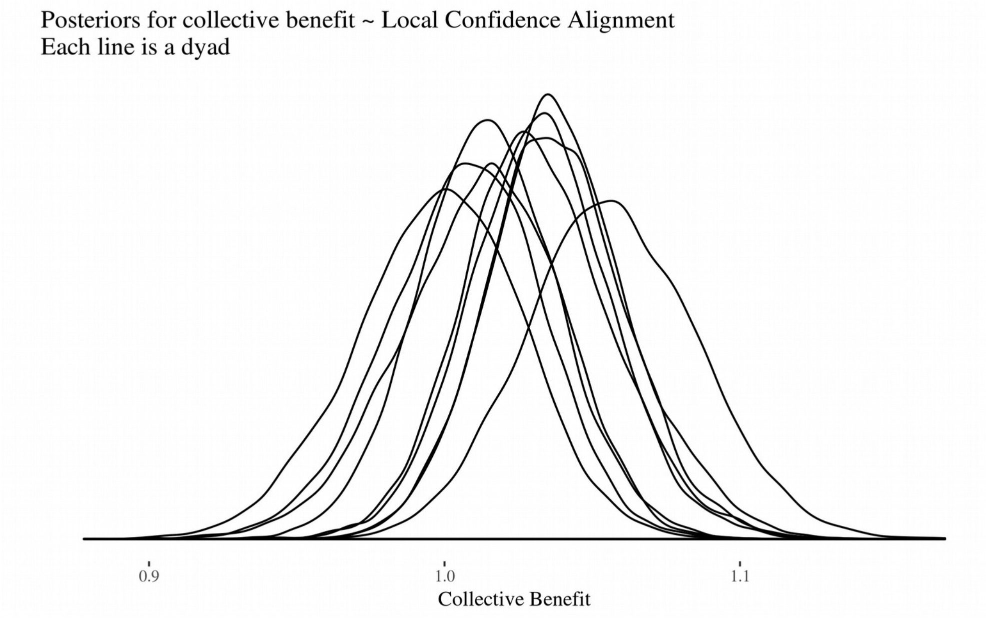

Local indiscriminate alignment
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
5
10
15
20
25
Intensity (dB)
Group
Individual


Cosine between word sets
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
Group
Individual
5
10
15
20
25
Intensity (dB)


Global confidence convergence
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
Group
Individual
5
10
15
20
25
Intensity (dB)


RQA: Alignment L
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
Group
Individual
5
10
15
20
25
Intensity (dB)


RQA: Alignment ENTR
Proportion correct answers
100%
75%
50%
25%
0%
Low (-3 SD)
Average
High (+3 SD)
Group
Individual
5
10
15
20
25
Intensity (dB)


Posteriors for collective benefit ~ Local Confidence Alignment
Each line is a dyad
0.9
1.0
1.1
Collective Benefit

Correlations between linguistic features

Especially the RQA features (bottom right quarter) appear highly correlated.

Stan code for the local confidence alignment model

```
data{
   // number of data points, groups, and vowels
   int<lower=1> N;
   int<lower=1> N_group;
   int<lower=1> N_vowel;

   // binomial data for the two psychometric functions
   int<lower=0> k_ind[N];
   int<lower=0> k_gro[N];
   int<lower=0> n_ind[N];
   int<lower=0> n_gro[N];

   // input data
   int group[N];
   int vowel[N];
   vector[N] intensity;

   // collective benefit model
   vector[N_group] local_confidence;
}

parameters{
   // random effects
   vector[N_group] ia_group;
   vector[N_group] ib_group;
   vector[N_group] ga_group;

   vector[N_vowel] ia_vowel;
   vector[N_vowel] ib_vowel;
   vector[N_vowel] ga_vowel;
   vector[N_vowel] gb_vowel;

   // main effects
   real ga;
   real gb;
   real ia;
   real ib;



   // parameters for the collective benefit model
   real<lower=0> collective_benefit[N_group];
   real<lower=0> collective_sigma;
```

```
    real col_a;
    real col_b;

}

transformed parameters{
    vector[N_group] gb_group;
    vector<lower=0,upper=1>[N] gb_group_theta;

    for (i in 1:N) {
        // collective benefit = inv_logit(dyad slope) / inv_logit(individual slope)
        // isolate for the group effect of the dyad slope
        gb_group_theta[i] = inv_logit(ib + ib_vowel[vowel[i]] + ib_group[group[i]]) * collective_benefit[group[i]];
        gb_group[group[i]] = logit(gb_group_theta[i]) - (gb + gb_vowel[vowel[i]]);
    }
}



model{
    // intermediate parameters for the psychometric curves
    vector[N] gtheta;
    vector[N] gA;
    vector[N] gB;

    vector[N] itheta;
    vector[N] iA;
    vector[N] iB;

    // intermediate parameter for the collective benefit model
    vector[N] collective_mu;

    // priors for the psychometric curves
    ia_group ~ normal(0,1);
    ib_group ~ normal(0,1);
    ga_group ~ normal(0,1);


  ia_vowel ~ normal(0,1);
    ib_vowel ~ normal(0,1);
    ga_vowel ~ normal(0,1);
    gb_vowel ~ normal(0,1);

    ia ~ normal(0, 1);
    ib ~ normal(0, 1);
```

```
  ga ~ normal(0, 1);
  gb ~ normal(0, 1);

  // psychometric function for both individuals and groups
  gA = ga + ga_vowel[vowel] + ga_group[group];
  gB = gb + gb_vowel[vowel] + gb_group[group];

  iA = ia + ia_vowel[vowel] + ia_group[group];
  iB = ib + ib_vowel[vowel] + ib_group[group];

  for (i in 1:N){
    // theta is not actually a rate parameter since I'm using binomial_logit
    gtheta[i] =  gA[i] + gB[i] * intensity[i];
    itheta[i] =  iA[i] + iB[i] * intensity[i];
  }

  k_gro ~ binomial_logit(n_gro, gtheta);
  k_ind ~ binomial_logit(n_ind, itheta);

  // priors for the collective benefit model
  col_a ~ normal(1,1);
  col_b ~ normal(0,1);
  collective_sigma ~ cauchy(0,2);

  // likelihood function for the collective benefit model
  for (k in 1:N_group) {
    collective_mu[k] = col_a + col_b * local_confidence[k];
    collective_benefit[k] ~ normal(collective_mu[k], collective_sigma);
  }
}
```