Description of Inter-Item Timings (IRTs)

Tyler Peckenpaugh

3/29/2019

Contents

Inter-item timing

Subjects were asked to read each sentence twice, once with no preview at all, and then again after unlimited preview. Inter-reading time (IRT) is a measure of the amount of time between when a subject stops speaking after a cold reading and when they begin speaking for a previewed reading.

> IRT = delay after the end of a cold reading and before the start of a previewed reading

Practically, this was done over 1548 recordings (33 participants, 48 items = 1584 pairs, with 36 missing data). This was measured using Google's WebRTC Voice Activity Detection (VAD) over .wav files that had been subjected to a high-pass filter with a low threshold of 0 to 500Hz[1] using the highest aggressiveness that yielded good results, depending on the noise level in the recording.

[1] a low hum in the room needed to be accounted for; the exact algorithm is available at github (URL: bit.ly/2uMrcrG)

Description and cleanup

The following section details the IRT data and the outlier removal and resulting participant attrition.

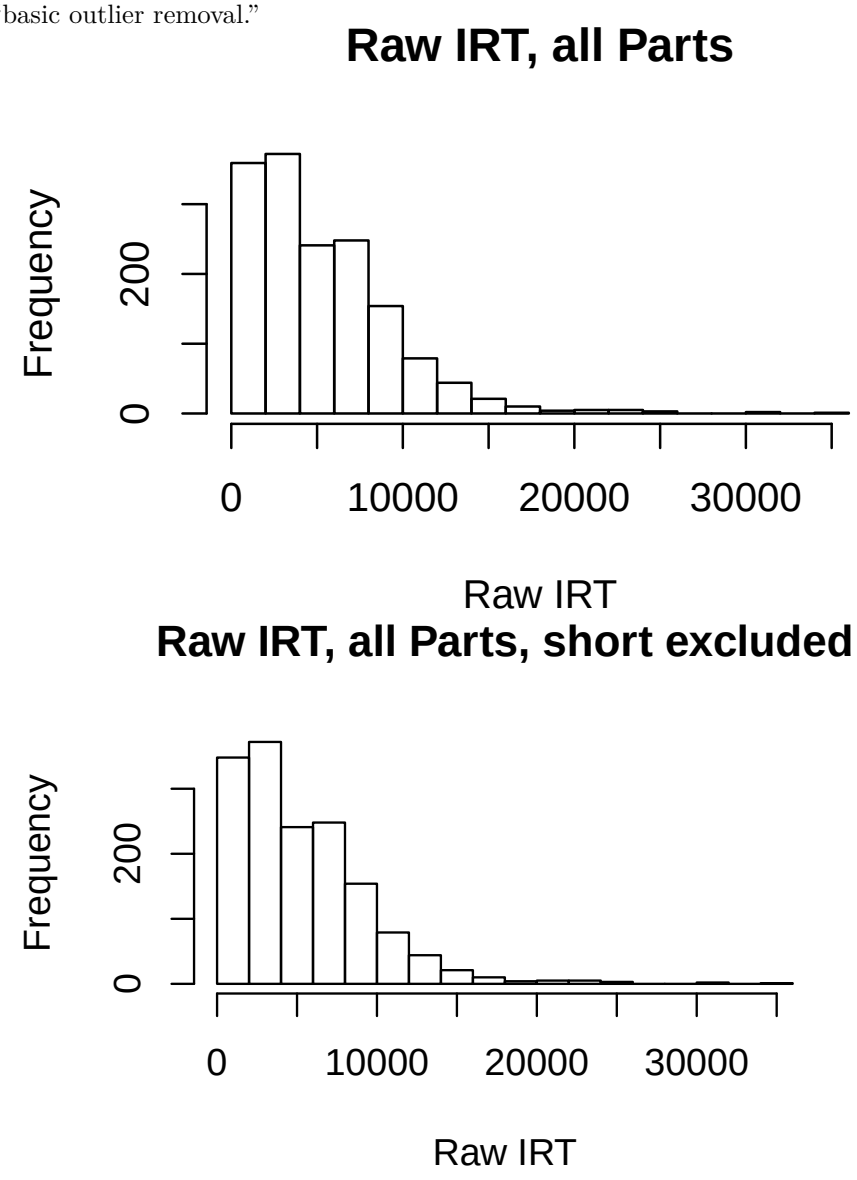
Distribution of IRTs, all participants

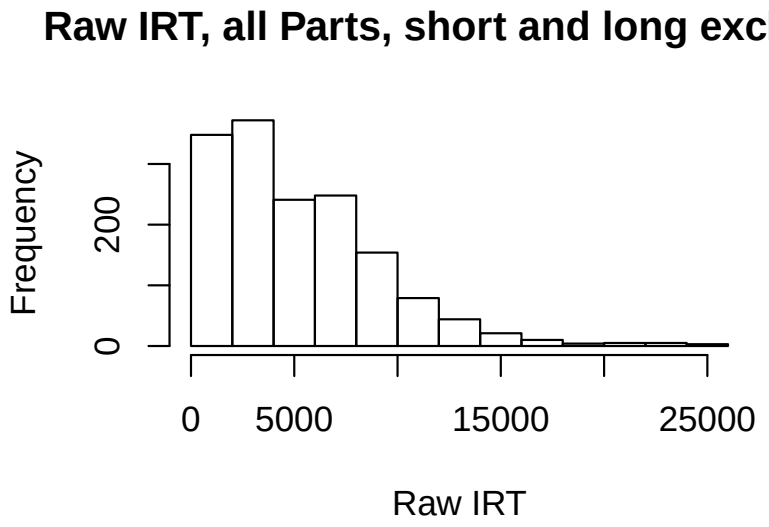
The overal mean IRT for all participants, all items (including fillers), and all conditions is 5317.99ms (sd = 4115.16). The highest IRT was 35762.85ms.

The following histograms show the distribution of IRT across all items and all participants. In the second graph, overly short IRTs (shorter than 200ms; 11[2] such data) are excluded. In the third, overly long (longer than 25s; 3 such data) and overly short IRTs are excluded.

[2] This is ??% of the 1548 total available data

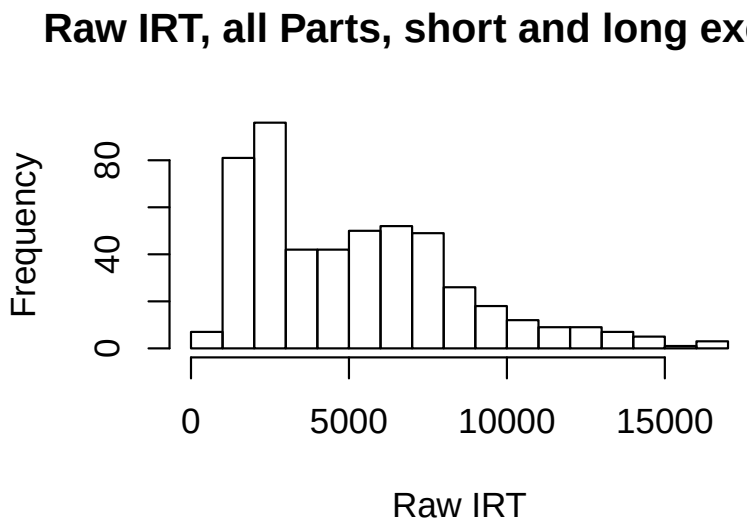
The third graph represent what I will call data that has undergone "basic outlier removal."





IRTs were finally winsorized to lessen the impact of outliers.

A question for DCB: should the IRTs be winsorized by a participant's mean/sd for all items (including fillers), or only by experimental item mean/sd? I assume the former in this document. Should this be done before or after basic outlier removal? I assume after in this document.



Missing data and attrition

Due to noise in recordings and/or technical difficulties during data collection, a number of IRTs are missing for experimental items in the data. The following table shows which participants are missing how many IRTs; ideally each would have 48 IRTs and 16 experimental IRTs.

Table 1: Missing data, by participant

	Missing IRTs	Available % of IRTs	Missing experimental IRTs	Available % of experimental IRTs
1	0	100%	0	100%
2	0	100%	0	100%
3	0	100%	0	100%
4	0	100%	0	100%
5	23	52.08%	8	50%
6	0	100%	0	100%
7	0	100%	0	100%
8	1	97.92%	0	100%
9	0	100%	0	100%
10	0	100%	0	100%
11	12	75%	5	68.75%
12	0	100%	0	100%
13	0	100%	0	100%
14	0	100%	3	81.25%
15	0	100%	0	100%
16	0	100%	0	100%
17	0	100%	0	100%
19	0	100%	2	87.5%
20	0	100%	0	100%
21	0	100%	0	100%
22	0	100%	0	100%
201	0	100%	0	100%
203	0	100%	0	100%
204	0	100%	1	93.75%
205	0	100%	0	100%
206	0	100%	0	100%
207	0	100%	0	100%
208	0	100%	0	100%
209	0	100%	0	100%
210	0	100%	0	100%
212	0	100%	0	100%
214	0	100%	0	100%
215	0	100%	0	100%

The 13 participants missing more than 1 experimental IRTs (5, 11, 14, 19) are excluded.

Subjects with overall mean IRTs that are very short (< 2200) or very long (> 10000) are also excluded (6, 13, 20, 201, 203, 204, 208)

Group sizes after attrition

The following table[3] shows how the participants are distributed across groupsafter attrition. Ideally, there would be 4 per group-order cell, but because of attrition the cells are uneven. Because regression is able to account for uneven groups, this defect will hopefully not play an important role in the analyses that follow.

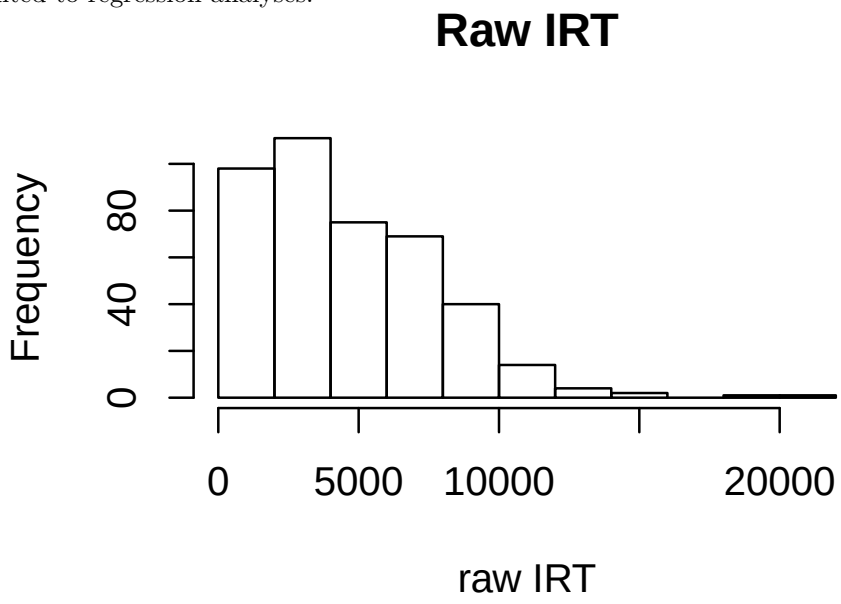
[3] There are 5 particpiants in Group 1, Split BA because I ran four participants per group-order, and then one extra who happened to be assigned to group 1, split BA; and by happenstance, none of the participants from that cell needed to be excluded.

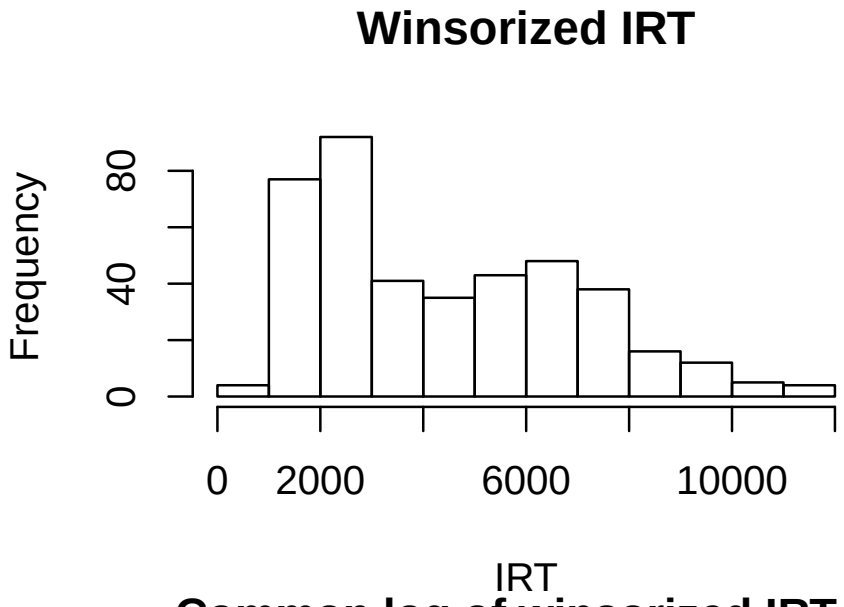
Table 2: Group/order totals after attrition

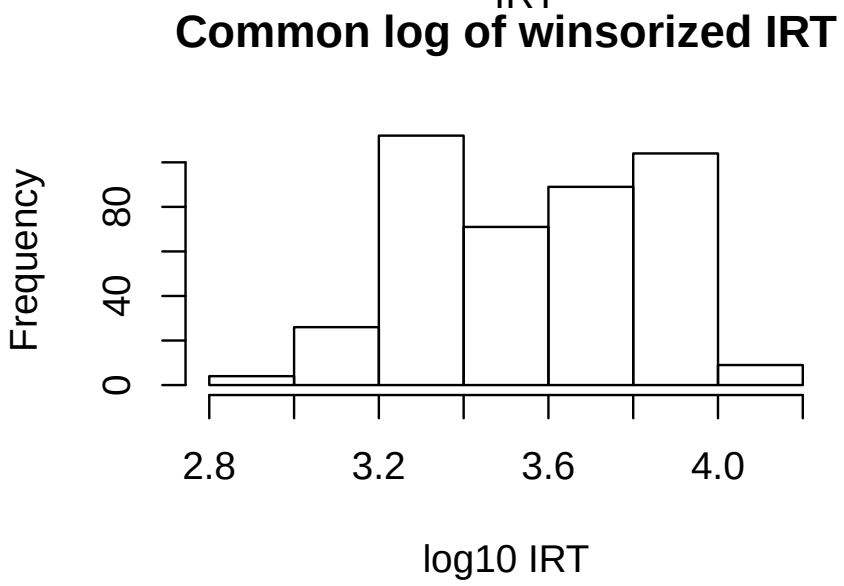
	Split AB	Split BA	Group Total
Group 1	3	5	8
Group 2	4	3	7
Group 3	2	3	5
Group 4	3	3	6
Split Total	12	14	26

Distribution of experimental item IRT after attrition

The following histograms show the distribution of experimental item IRTs after attrition, and then the Winsorized IRTs, and finally the common log of winsorized IRTs, which are the shape of the data most suited to regression analyses.







Mean and SD of winsorized IRT by condition

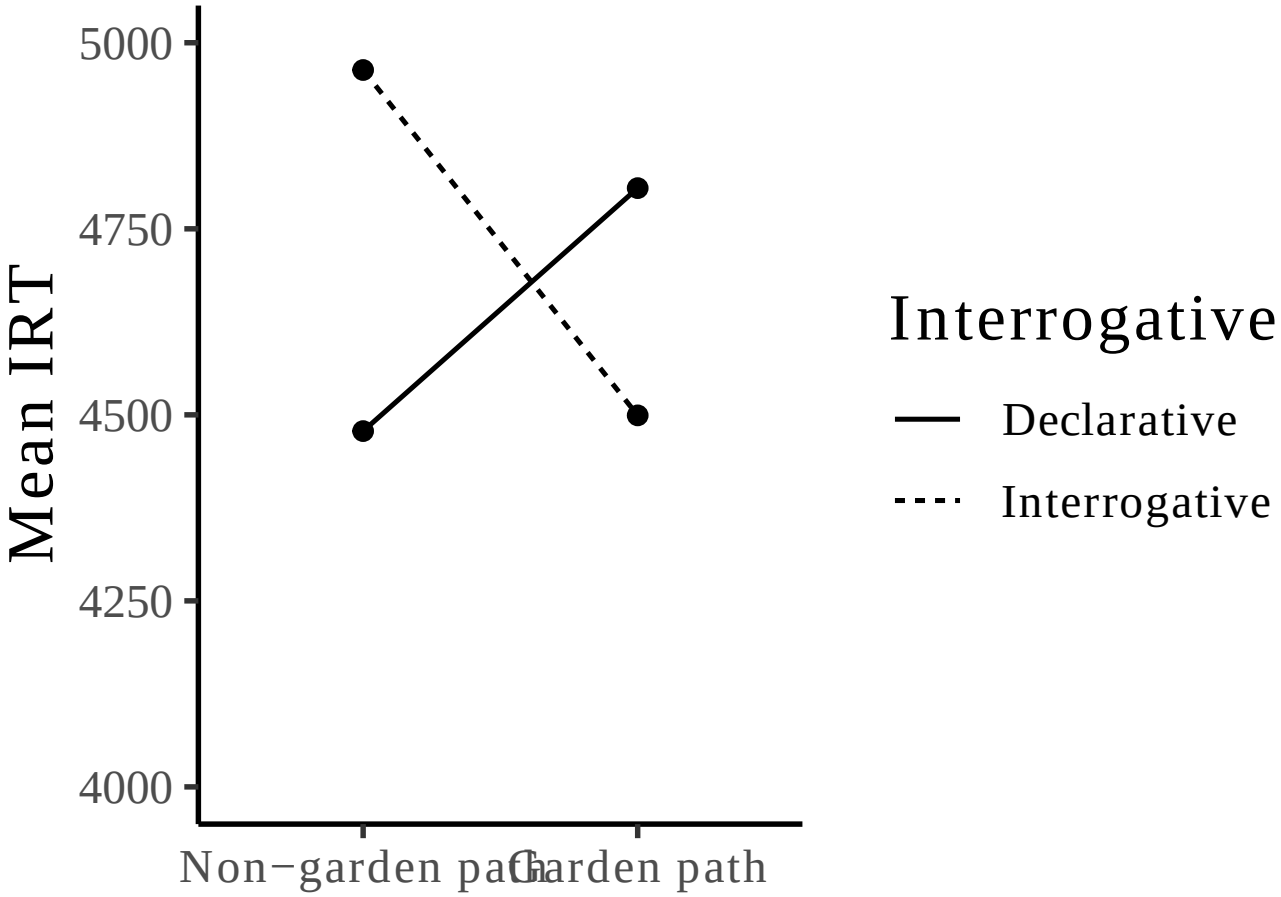
If we assume that interrogative PP-attachment garden paths are easier to process as an interrogative than in the declarative, and that IRT represents how difficult a sentence is to process, we would expect the difference in mean IRT to be larger for declarative garden paths compared to declarative controls than for the same comparison of interrogatives.

The means of the Winsorized IRT by condition indeed show this pattern.

Table 3: Mean experimental IRT by condition

	Garden path	Non-garden path
Declarative	4804.662	4478.158
Interrogative	4499.242	4963.408

Boop



The difference in mean IRT acriss ± for declaratives is 326.5033654; for interrogatives, it's -464.1650728. This is a difference of 790.6684382, representing the impact of ±GP for +Q compared to -Q. This supports the hypothesis that garden paths are easier to comprehend when presented as interrogative. It is strange that the garden-path interrogatives appear to be comprehended more quickly than the non-garden path interrogatives. A possible explanation will be explored in the discussion section.

Item and subject variation

There is variation across participants in terms of whether or not they show this pattern.

Number of participants who show predicted pattern

In the analyzed data,16 of 26 participants show the expected pattern.

Table 4: Mean IRT by condition and participant

participant	-Q -GP	-Q +GP	+Q -GP	+Q +GP	pattern
1	2321.508	2607.383	2713.935	2573.972	TRUE
2	5952.547	6220.012	5214.915	6914.200	TRUE
3	5940.587	6621.430	6401.395	5796.435	TRUE
4	7753.830	6426.230	8993.090	7364.613	FALSE
7	5338.530	5580.860	6119.302	6089.198	TRUE
8	5485.828	4944.953	2931.530	5922.315	FALSE
9	6167.350	6326.610	6236.807	6210.695	TRUE
10	5064.172	5066.130	5376.845	5532.590	TRUE
12	7396.735	7284.235	8725.215	8569.370	FALSE
15	5000.038	3622.005	5294.510	4421.972	FALSE
16	2203.932	2113.835	2252.343	2164.775	FALSE
17	2590.157	4947.785	4880.695	2641.673	TRUE
20	10217.142	9322.782	11525.555	9796.060	FALSE
21	5091.373	5658.535	5757.895	5662.045	TRUE
201	1463.557	1531.938	1330.318	1543.527	TRUE
203	1605.760	1733.925	1583.150	1812.240	TRUE
204	2052.124	1729.273	2021.573	1719.567	FALSE
205	5467.005	7125.280	5642.820	3413.762	TRUE
206	3715.318	3407.920	3291.235	3073.832	FALSE
207	4545.120	4991.815	4846.675	3648.440	TRUE
208	1788.898	2283.023	1930.420	2261.302	TRUE
209	2214.637	3042.070	2158.830	2522.142	TRUE
210	3897.847	4643.092	5679.913	2205.215	TRUE
212	3456.225	3264.745	3453.795	4052.240	FALSE
214	2986.100	2776.610	2994.983	1644.010	FALSE
215	3029.610	3712.920	3590.963	3689.343	TRUE

Number of items that show predicted pattern

For items, 7 of 16 show the pattern.

Table 5: Mean IRT by condition and item

item	-Q -GP	-Q +GP	+Q -GP	+Q +GP	pattern
1	4351.327	4042.900	5017.644	2860.016	FALSE
2	5925.486	5045.800	3542.836	4611.902	FALSE
3	4924.914	3718.398	4390.247	5372.898	FALSE
4	4340.826	5082.329	5538.217	3156.350	TRUE
5	4383.834	4295.718	3677.701	3290.423	FALSE
6	4291.746	4320.389	4552.698	3967.449	TRUE
7	3481.002	5220.940	5153.560	4969.779	TRUE
8	4869.698	4486.814	4786.232	5515.082	FALSE
9	3477.217	5741.277	5531.248	4399.666	TRUE
10	4943.123	4631.096	4801.387	3696.528	FALSE
11	3487.558	3930.710	4897.048	5527.831	TRUE
12	3918.147	3858.007	5466.719	3260.450	FALSE
13	5134.838	4212.736	3209.472	5388.462	FALSE
14	3621.984	4666.264	4503.277	3236.543	TRUE
15	2719.359	3866.453	4901.150	4975.746	TRUE
16	5038.232	4615.556	3963.138	3678.411	FALSE

Analyses

The following models explore the effect of garden path (±GP) and interrogativeness (±Q) on IRT.

Regression analyses

Regression models with fixed effects of ±GP and ±Q were run, one including the interaction of ±GP and ±Q and one without the interaction term. Both included random effects for item and participant.

Models with random slopes for GP, Q, and their interaction for both error terms fails to converge. A model with random slopes for just GP and Q ain effects likewise fails to converge. Models without random slopes of fixed effects were used.

The interaction model represents a better fit; the non-interaction model represents a singular fit that is worse overall ($X^2 = 4.836$, $p < 0.03$). This supports the hypothesis and the earlier observation over the means that garden paths are more difficult as declaratives than interrogatives.

The relevance of random effects were also tested, by comparing models that exclude each to the model with both random effects (I call this the "full model" in what follows).

Removing the random effect of item does not degrade the model in a stastically significant way (AIC~full model~ = -305; AIC~no item

Table 6: Mixed effects model

	Dependent variable:	
	Common log of IRT	
	(1)	(2)
Garden path	0.026	−0.008
	(0.022)	(0.016)
Interrogative	0.029	−0.005
	(0.022)	(0.016)
Interaction	−0.069**	
	(0.031)	
Constant	3.559***	3.576***
	(0.045)	(0.045)
Observations	415	415
Log Likelihood	130.600	128.182
Akaike Inf. Crit.	−247.199	−244.364
Bayesian Inf. Crit.	−219.001	−220.194
Note:	*p<0.1; **p<0.05; ***p<0.01	

error~ = -307; -X^2 = 0.17, p = 0.68), but removing the random effect of participant does (AIC~full model~ = -305; AIC~no participant error~ = 47; X^2 = 348.24, p = 0). The model with no random effects is worse than both the full model and the model with only item removed. Ultimately, it's difficult to select between the full model and the "no item" model, as both offer strong fits with more or less the same outcome.