Online supplement to: "How to Neutralize a Finite Clause Boundary: Phase Theory and the Grammar of Bound Pronouns" (Grano and Lasnik 2018, *LI* 49)

Appendix: Statistical analysis of experimental results

In this appendix, we describe in greater detail the statistical analysis of the experimental investigation described in section 3.4 of the main article. In selecting tests for statistical analysis, we assume following Sprouse (2011) and others that sentence acceptability judgments do not necessarily conform to a ratio scale; that is, we assume that participants treat the seven points on the Likert scale as defining a ranking, but we do not assume that the difference between a rating of 2 and a rating of 3, for example, is the same as the difference between a rating of 3 and a rating of 4. This means that the resulting data have to be treated as ordinal data rather than as ratio-scale data.

The input to the statistical analysis for Experiment 1 is 2,475 test sentences rated on a scale of 1 to 7. We treat the rating as the dependent variable. Each of the 2,475 sentences is coded for two factors that constitute the independent variables. The *phenomenon* factor consists of the three categories listed in (A1a) and the *condition* factor consists of the six categories listed in (A1b).

(A1) a. Phenomenon: Comparative deletion, Multiple questions, *Too/Enough* movement
b. Condition: BaseLine, NonFinite, BoundSubj, BoundObj, BoundPoss, NoBinding

The first test we employ is an Independent-Samples Kruskal-Wallis Test, a rank-based nonparametric test similar to a one-way ANOVA but appropriate for ordinal (non-ratio-scale) data (see Sheskin 2003). This test allows us to determine whether or not the distribution of sentence ratings is the same across the different categories of a chosen factor. To run this test and all the other statistical tests described in what follows, we use IBM SPSS Statistics Version 24.

Applied to the phenomenon factor, the Kruskal-Wallis Test indicates that the distribution of ratings is *not* the same across the different categories of the phenomenon factor ($X^2(2) = 107.130$, $p < 0.01$). Furthermore, pairwise comparisons reveal that each phenomenon gives rise to a rating profile that is significantly different from each other phenomenon. These pairwise comparisons are shown in Table A1, with significance values adjusted by the Bonferroni correction for multiple tests. Taken together with the mean rank for each phenomenon indicated in Table A2, this analysis supports the conclusion that the three phenomena investigated in the experiment conform to the acceptability cline in (A2): on the whole, comparative deletion sentences (Mean Rank = 1422.17) were rated higher ($p < 0.01$) than *too/enough* movement sentences (Mean Rank = 1229.29), which were in turn rated higher ($p < 0.01$) than multiple questions (Mean Rank = 1062.54). While interesting and worthy of further study, we take this result to be orthogonal to our main purpose, which is to establish how ratings vary as a function of the condition factor. That being said, the cline that emerges here may bear an interesting cross-linguistic connection to Rizzi's (1982) claim that Italian does not allow multiple questions.

(A2) Comparative deletion > *Too/Enough* movement > Multiple questions

Sample1 - Sample2	Test Statistic	Standard Error	Standard Test Statistic	Significance	Adjusted Significance
Multiple Questions – *too/enough* Movement	-166.750	34.776	-4.795	.000	.000***
Multiple Questions - Comparative Deletion	359.627	34.776	10.341	.000	.000***
too/enough Movement - Comparative Deletion	192.877	34.776	5.546	.000	.000***

Table A1: Experiment 1 pairwise comparisons of phenomena

Phenomenon	Mean Rank
Comparative Deletion	1,422.17
too/enough Movement	1,229.29
Multiple Questions	1,062.54

Table A2: Experiment 1 mean ranks for phenomena

Applied to the condition factor, the Kruskal-Wallis Test indicates that the distribution of ratings is *not* the same across the different conditions ($X^2(5) = 325.701$, $p < 0.01$). Pairwise comparisons reveal that each condition gives rise to a rating profile significantly different from each other condition ($p < 0.01$), except for the BoundPoss, BoundObj, and NoBinding conditions which are not significantly different from one another ($p = 1$). These pairwise comparisons are shown in Table A3. Taken together with the mean ranks for each condition (Table A4), this analysis supports the conclusion that the six conditions investigated in Experiment 1 conform to the cline of acceptability indicated in (A3): BaseLine sentences were rated as most acceptable (Mean Rank = 1841.69), followed by sentences with a nonfinite embedded clause (Mean Rank = 1494.58), followed by sentences with a finite embedded clause containing a bound pronominal subject (Mean Rank = 1258.58). At the low end are sentences with an embedded finite clause containing a bound pronominal object (Mean Rank = 1064.88), a bound subject-internal possessor (Mean Rank = 1024.02), or no bound pronoun (Mean Rank = 1046.09). These three give rise to ratings not significantly different from one another.

(A3) BaseLine > NonFinite > BoundSubject > {BoundObj = NoBinding = BoundPoss}

Sample1 - Sample2	Test Statistic	Standard Error	Standard Test Statistic	Significance	Adjusted Significance
BoundPoss-NoBinding	-22.074	47.087	-.469	.639	1
BoundPoss -BoundObj	40.861	47.087	.868	.386	1
BoundPoss -BoundSubj	-234.564	47.087	-4.982	.000	.000***
BoundPoss -NonFinite	-470.560	47.087	-9.993	.000	.000***
BoundPoss -BaseLine	817.672	57.669	14.179	.000	.000***
NoBinding - BoundObj	18.787	47.087	.399	.690	1
NoBinding - BoundSubj	212.490	47.087	4.513	.000	.000***
NoBinding - NonFinite	-448.486	47.087	-9.525	.000	.000***
NoBinding - BaseLine	795.598	57.669	13.796	.000	.000***
BoundObj -BoundSubj	-193.703	47.087	-4.114	.000	.000***
BoundObj - NonFinite	-429.699	47.087	-9.126	.000	.000***
BoundObj - BaseLine	776.811	57.669	13.470	.000	.000***
BoundSubj - NonFinite	-235.996	47.087	-5.012	.000	.000***
BoundSubj - BaseLine	583.108	57.669	10.111	.000	.000***
NonFinite - BaseLine	347.112	57.669	6.019	.000	.000***

Table A3: Experiment 1 pairwise comparisons of conditions

Condition	Mean Rank
BaseLine	1,841.69
NonFinite	1,494.58
BoundSubj	1,258.58
BoundObj	1,064.88
NoBinding	1,046.09
BoundPoss	1,024.02

Table A4: Experiment 1 mean ranks for conditions

A limitation of the Kruskal-Wallis Test is that it only allows us to test one factor at a time: phenomenon or condition. To remedy this, we employ a more powerful statistical technique: a Generalized Estimating Equations (GEE) analysis. GEE is a technique appropriate for ordinal data with multiple independent variables, similar to a generalized multiple linear regression but different in that it requires fewer assumptions about the data and it models population averages rather than yielding subject-specific estimates (see e.g. Kenward, Lesaffre and Molenberghs 1994 for a discussion of GEE in the context of a psychiatric study). Applied to the data in Experiment 1, GEE yields the results indicated in Table A5.

Of most relevance to us are the rows labeled A-C and the rows labeled D-I, respectively. Looking first at the rows labeled A-C, the *too/enough* Movement category in row C is (arbitrarily) selected as a baseline, and the B column shows the increase in log odds for the other categories in this factor, namely Multiple Questions and Comparative Deletion, yielding a rating that is higher than the rating for a *too/enough* Movement sentence. The Exp(B) column translates this figure into an odds ratio: odds ratios that are greater than 1 indicate an increased likelihood of a higher rating whereas ratios less than 1 indicate a decreased likelihood of a higher rating. Hence, we see confirmation of the conclusion from the pairwise comparisons that ratings for *too/enough* Movement sentences are significantly *higher* in odds ratio than ratings for Multiple

Questions sentences (Exp(B) = 0.618, $p < 0.01$) and significantly *lower* in odds ratio (0.62) than ratings for Comparative Deletion sentences (Exp(B) = 1.570, $p < 0.01$).

Turning to the rows labeled D-I, the NoBinding condition is (arbitrarily) selected as a baseline, and the Exp(B) column indicates the odds ratio for each of the other conditions in yielding a rating that is higher than that for NoBinding. We see, also consistent with the pairwise comparisons shown above, that the odds ratios for the NoBinding sentences are not significantly different from those for BoundObj (Exp(B) = 1.016, $p = 0.917$) or BoundPoss sentences (Exp(B) = 0.959, $p = 0.786$), but are significantly *lower* than those for BoundSubj (Exp(B) = 1.680, $p = 0.002$), NonFinite (Exp(B) = 3.272, $p < 0.01$), and BaseLine (Exp(B) = 9.608, $p < 0.01$) sentences.

Parameter Estimates

Parameter		B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
Threshold	[Choice_DV=1]	-2.041	0.1520	-2.339	-1.743	180.362	1	0.000	0.130	0.096	0.175
	[Choice_DV=2]	-0.939	0.1350	-1.203	-0.674	48.296	1	0.000	0.391	0.300	0.510
	[Choice_DV=3]	-0.075	0.1382	-0.346	0.196	0.293	1	0.588	0.928	0.708	1.217
	[Choice_DV=4]	0.516	0.1363	0.249	0.783	14.317	1	0.000	1.675	1.282	2.188
	[Choice_DV=5]	1.375	0.1456	1.090	1.661	89.291	1	0.000	3.957	2.975	5.263
	[Choice_DV=6]	2.666	0.1574	2.358	2.975	286.817	1	0.000	14.385	10.566	19.584
A. [Multiple Questions]		-0.482	0.1118	-0.701	-0.263	18.553	1	0.000	0.618	0.496	0.769
B. [Comparative Deletion]		0.451	0.1202	0.215	0.686	14.066	1	0.000	1.570	1.240	1.987
C. [*too/enough* Movement]		0[a]							1		
D. [BaseLine]		2.263	0.2115	1.848	2.677	114.497	1	0.000	9.608	6.348	14.543
E. [NonFinite]		1.185	0.1601	0.872	1.499	54.832	1	0.000	3.272	2.391	4.477
F. [BoundSubj]		0.519	0.1644	0.196	0.841	9.950	1	0.002	1.680	1.217	2.319
G. [BoundObj]		0.016	0.1532	-0.284	0.316	0.011	1	0.917	1.016	0.753	1.372
H. [BoundPoss]		-0.042	0.1548	-0.346	0.261	0.074	1	0.786	0.959	0.708	1.299
I. [NoBinding]		0[a]							1		
(Scale)		1									

Dependent Variable: Choice_DV
Model: (Threshold), Phenomenon, Condition
a. Set to zero because this parameter is redundant.

Table A5: Experiment 1 Generalized Estimating Equation Parameter Estimates

We now turn our attention to the analysis of the data in Experiment 2. Since Experiment 2 is identical in setup to Experiment 1 except that the sentences instantiating the BoundObj and BoundPoss conditions are replaced by sentences that instantiate 1pSubj and 2Subj conditions, respectively, we employ the same statistical tests. As expected, the Kruskal-Wallis Test applied to the phenomenon factor in the Experiment 2 data indicates that the distribution of ratings is *not* the same across the different categories of the phenomenon factor ($X^2(2) = 86.409, p < 0.01$). As shown in Tables A6 and A7, we see the same cline of acceptability schematized in (A2) as we did for the Experiment 1 data. Also as expected, the Kruskal-Wallis Test applied to the condition factor indicates that the distribution of ratings is *not* the same across the different conditions ($X^2(5) = 349.406, p < 0.01$). The pairwise comparisons and mean ranks are shown in Tables A8-

A9. Taken together, they support the conclusion that the sentences tested in Experiment 2 conform to the cline of acceptability schematized in (A4). Of particular interest is the observation that the 1pSubj and 2pSubj conditions give rise to rating profiles that are not significantly different from that of the NoBinding condition.

(A4) BaseLine > NonFinite > BoundSubject > {1pSubj = 2pSubj = NoBinding}

Sample1 - Sample2	Test Statistic	Standard Error	Standard Test Statistic	Significance	Adjusted Significance
Multiple Questions – *too/enough* Movement	-197.550	34.744	-5.686	0.000	0.000***
Multiple Questions - Comparative Deletion	320.315	34.775	9.211	0.000	0.000***
too/enough Movement - Comparative Deletion	122.764	34.775	3.530	0.000	0.001***

Table A6: Experiment 2 pairwise comparisons of phenomena

Phenomenon	Mean Rank
Comparative Deletion	1,384.37
too/enough Movement	1,261.61
Multiple Questions	1,064.06

Table A7: Experiment 1 mean ranks for phenomena

Sample1 - Sample2	Test Statistic	Standard Error	Standard Test Statistic	Significance	Adjusted Significance
2pSubj-1pSubj	38.236	57.122	.811	.417	1
2pSubj-NoBinding	61.978	47.096	-1.316	.188	1
2pSubj-BoundSubj	-292.945	47.096	-6.220	.000	.000***
2pSubj-NonFinite	-542.577	47.096	-11.521	.000	.000***
2pSubj-BaseLine	-819.462	57.659	-14.056	.000	.000***
1pSubj-NoBinding	-23.742	47.070	-.504	.614	1
1pSubj-BoundSubj	-254.709	47.070	-5.411	.000	.000***
1pSubj-NonFinite	-504.341	47.070	-10.715	.000	.000***
1pSubj-BaseLine	-772.225	57.637	-13.398	.000	.000***
NoBinding-BoundSubj	230.967	47.043	4.910	.000	.000***
NoBinding-NonFinite	-480.599	47.043	-10.216	.000	.000***
NoBinding-BaseLine	748.483	57.616	12.991	.000	.000***
BoundSubj-NonFinite	-249.632	47.043	-5.306	.000	.000***
BoundSubj-BaseLine	517.517	57.616	8.982	.000	.000***
Nonfinite-BaseLine	267.884	57.616	4.649	.000	.000***

Table A8: Experiment 2 pairwise comparisons of conditions

Condition	Mean Rank
BaseLine	1,802.87
NonFinite	1,534.98
BoundSubj	1,285.35
NoBinding	1,054.39
1pSubj	1,030.64
2pSubj	992.41

Table A9: Experiment 2 mean ranks for conditions

Finally, the results of the GEE analysis as applied to the data from Experiment 2 are as indicated in Table A10. Here we see results that are consistent with the conclusions from the Kruskal-Wallis test. As seen in rows A-C, *too/enough* Movement sentences are rated significantly *higher* than Multiple Questions (Exp(B) = 0.567, $p < 0.01$) but *lower* than Comparative Deletion sentences in a way that trends toward significance (Exp(B) = 1.322, $p = 0.014$). As seen in rows D-I, ratings for NoBinding sentences are not significantly different than those for 2pSubj sentences (Exp(B) = 0.848, $p = 0.302$) or 1pSubj sentences (Exp(B) = 0.919, $p = 0.590$), but significantly lower than those for BoundSubj sentences (Exp(B) = 1.773, $p < 0.01$), NonFinite sentences (Exp(B) = 3.334, $p < 0.01$), and BaseLine sentences (Exp(B) = 8.405, $p < 0.01$).

Parameter Estimates

Parameter		B	Std. Error	95% Wald Confidence Interval Lower	95% Wald Confidence Interval Upper	Hypothesis Test Wald Chi-Square	df	Sig.	Exp(B)	95% Wald Confidence Interval for Exp(B) Lower	95% Wald Confidence Interval for Exp(B) Upper
Threshold	[Choice_DV=1]	-2.076	0.1527	-2.375	-1.777	184.771	1	0.000	0.125	0.093	0.169
	[Choice_DV=2]	-0.933	0.1357	-1.199	-0.667	47.339	1	0.000	0.393	0.301	0.513
	[Choice_DV=3]	-0.099	0.1338	-0.361	0.164	0.545	1	0.460	0.906	0.697	1.178
	[Choice_DV=4]	0.615	0.1359	0.348	0.881	20.455	1	0.000	1.849	1.417	2.413
	[Choice_DV=5]	1.493	0.1418	1.216	1.771	110.921	1	0.000	4.453	3.372	5.879
	[Choice_DV=6]	2.610	0.1597	2.297	2.923	267.078	1	0.000	13.602	9.946	18.602
A. [Multiple Questions]		-0.568	0.1207	-0.804	-0.331	22.096	1	0.000	0.567	0.447	0.718
B. [Comparative Deletion]		0.279	0.1138	0.056	0.502	6.001	1	0.014	1.322	1.057	1.652
C. [*too/enough* Movement]		0[a]							1		
D. [BaseLine]		2.129	0.2434	1.652	2.606	76.489	1	0.000	8.405	5.216	13.544
E. [NonFinite]		1.204	0.1409	0.928	1.480	73.029	1	0.000	3.334	2.529	4.394
F. [BoundSubj]		0.572	0.1374	0.303	0.842	17.356	1	0.000	1.773	1.354	2.320
G. [1pSubj]		-0.085	0.1572	-0.393	0.223	0.291	1	0.590	0.919	0.675	1.250
H. [2pSubj]		-0.164	0.1594	-0.477	0.148	1.064	1	0.302	0.848	0.621	1.160
I. [NoBinding]		0[a]							1		
(Scale)		1									

Dependent Variable: Choice_DV
Model: (Threshold), Phenomenon, Condition
a. Set to zero because this parameter is redundant.

Table A10: Experiment 2 Generalized Estimating Equation Parameter Estimates

References

Kenward, Michael G., Emmanuel Lesaffre, and Geert Molenberghs. 1994. An application of maximum likelihood and generalized estimating equations to the analysis of ordinal data from a longitudinal study with cases missing at random. *Biometrics* 50:945–953.

Rizzi, Luigi. 1982. Violations of the wh-island constraint and the subjacency condition. In *Issues in Italian Syntax*, ed. L. Rizzi, 49–76. Dordrecht: Foris.

Sheskin, David. 2003. *Handbook of parametric and nonparametric statistical procedures*. Boca Raton, FL: Chapman & Hall/CRC.

Sprouse, Jon. 2011. A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language* 87:274–288.