Which sentences do speakers favor? ROC analysis of d-linking in filler-gap integration

In filler-gap dependencies (FGD), filler phrases with lexical content (*d-linked*) are sometimes reported to ease integration [1; cf. 2-3] and improve the acceptability of long-distance FGDs [4]. One plausible reason is that *d*-linking makes the filler easier to retrieve from memory at the gap-site, thus increasing the probability of successfully integrating the filler at the gap [1,4]. However, the increased offline acceptability reported for d-linking could in principle come from diverse underlying sources, which are conflated in standard acceptability judgment methodology. We address this issue using a Signal Detection Theoretic analysis of acceptability ratings. We ask whether the ameliorative effect of d-linking specifically improves a rater's ability to parse filler gap dependencies, as predicted by the retrieval hypothesis.

Design. Following [5], we reasoned that if *d*-linking improves the parser's ability to retrieve a filler, it should selectively improve sentence acceptability when filler retrieval is necessary to parse a long-distance FGD in a grammatical sentence, as in (1a/b). We constructed lexically-matched ungrammatical controls (1c/d) that did not require filler retrieval to process the same verb; so any effects of *d*-linking that are unrelated to retrieval should obtain in ungrammatical controls. We designed 32 items in 4 conditions and combined them with 64 fillers carefully designed to eliminate any superficial response strategies. We collected ratings on these sentences in three parallel word-by-word RSVP rating experiments. Each differed only in the response method: a binary 'yes'-'no' decision, followed by a three-point confidence rating (**Y/N+3-pt;** N_{subj} = 64); a numerical Likert-like scale (**Likert;** N_{subj} = 45); or a Likert-like scale labeled with confidence ratings (**Pseudo-Likert;** N_{subj} = 45). This methodological comparison served the additional goal of exploring how task demands may affect recovery of the latent variables that support an acceptability judgment.

Analysis. Data from each judgment method were analyzed two ways: (i) We constructed empirical ROC curves by scaling acceptable filler-gap dependencies against their unacceptable baselines [6]. From the ROCs we calculated d_a, a bias-free index of sensitivity to grammaticality; and *s*, or slope, which measures the relative variance of the grammatical/ungrammatical evidence distributions. (ii) We estimated ordinal mixed-effects regression models, parameterized to implement an unequal-variance signal detection analysis [7]. For each analysis, the retrieval hypothesis predicts greater sensitivity to grammaticality in *d*-linking conditions. Results are reported in Tables 1 and Tables 2.

Results. We failed to observe an effect of *d*-linking on sensitivity, in any method. Thus, we saw no indication that *d*-linking increased retrievability of the filler. There was, however, some indication from decision criterion placement (unreported) that *d*-linking influenced response bias instead [cf. 2]. We also consistently found that the variance in the evidence distribution for grammatical sentences was higher than the distribution for ungrammaticals; and in all tasks, *d*-linking numerically shrunk this variance (but significantly so only in Y/N+3-pt).

Discussion. We found that *d*-linking does not improve comprehenders' abilities to discriminate grammatical sentences from ungrammatical ones. This finding casts doubt on the hypothesis that d-linking improves retrieval success. We conjecture that it may lead to a less variable distribution of parsing outcomes. Following [2], this may stem from differences in how the reference set implied by the filler is accommodated, a process independent of filler retrieval at the gap.

Design

(1) Gram. (✓/*) x WhP (Bare/**D-linked)**

a/b. ✓ Who/**Which diva** do you believe that the parrot could imitate _?

c/d. * Who/**Which diva** _ believes that the parrot could imitate?

Ungram. c/d resolve filler-gap dependency early, but embedded V's transitivity is not satisfied.

Table 1 ROC Analysis: d_a and s

	Y/N + 3-pt confidence		**Pseudo-Likert**		**Likert**	
	d_a	s	d_a	s	d_a	s
bare	0.75 [.56 - .92]	0.77 [.65 - .89]	0.92 [.78 - 1.1]	0.91 [.76 - 1.1]	0.85 [.70 - 1.0]	0.81 [.70 - .96]
dlink	0.81 [.65 - .95]	0.92 [.81 - 1.1]	0.85 [.70 - .97]	0.83 [.69 - .99]	0.89 [.70 - 1.1]	0.82 [.69 - .97]

[95% confidence intervals] derived from bootstrap (R=1000) sampling.

Table 2 Results of mixed-effects ordinal regression

	Y-N + conf.	**Pseudo-Likert**	**Likert**
LOCATION EFFECTS (sensitivity)			
Gram:ungram	**-0.60 (.07)**	**-0.75 (.08)**	**-0.84 (.09)**
WhP:dlink	-0.04 (.07)	-0.07 (.08)	0.03 (.08)
Gram:ungram::WhP:dlink	0.06 (.08)	-0.00 (.10)	0.11 (.11)
SCALE EFFECTS (relative variance)			
Gramm:ungram	**-0.54 (.10)**	**-0.28 (.09)**	**-0.27 (.09)**
WhP:dlink	**-0.36 (.08)**	-0.13 (.08)	-0.07 (.09)
Gram:ungram::WhP:dlink	0.20 (.11)	0.02 (.11)	-0.04 (.12)

Bold coefficients have z-scores > 2. Location effects on probit scale. Variance ratios expressed on natural log scale. Random slopes included for location effects.

References. [1] Hofmeister, 2011. *Language and Cognitive Processes.* [2] Donkers et al. 2013. *Language and Cognitive Processes.* [3] Kaan et al., 2000. *Language and Cognitive Processes*. [4] Goodall, 2015. *Frontiers.* [5] McElree et al., 2003. *Journal of Memory and Language.* [6] MacMillan & Creelman, 2004. [7] Knoblauch & Maloney, 2012.