Supplementary Materials (Yang, Chan, Chang, & Kidd)

S1. Additional analyses of corpus data.

An anonymous reviewer rightly pointed out that one additional cue that influences RC acquisition and processing is head noun animacy (e.g., Kidd, Brandt, Lieven, & Tomasello, 2007; Mak, Vonk, & Schriefers, 2002). While animacy was not the focus of our research, we report here an analysis of noun animacy in a subset of our corpus data. Specifically, we coded both the head noun and the non-head NP for animacy (for RCs, the *RC-internal NP*, for RC-like structures, *the non-head NP*) in Mandarin adult child-directed speech from the following six corpora: AcadLang corpus (Zhou doi:10.21415/T5SC9D), Chang1 & Chang 2 corpus (Chang, 1998), Tong corpus (Deng & Yip, 2018) and Zhou 1 (Zhou, 2001) & Zhou 2 (Li & Zhou, 2004).

In Mandarin the head noun can be either expressed or null. Here we only analyse the animacy of expressed nouns since the animacy of unexpressed nouns is often difficult to determine from written transcripts. Table S1.1 reports the distribution of animate and inanimate nouns in general RC-like sequences; Table S1.2 reports the same distributions for genuine RC constructions only.

General RC-like sequences

Table S1.1
Animacy of the head and non-head NPs in subject and object RC-like sequences for DE and DCL construction types

		Non-head NP (animate)		**Non-head NP (inanimate)**	
		Animate Head	**Inanimate Head**	**Animate Head**	**Inanimate Head**
DE	**SRC-like**	3.7% (21/567)	42.0% (238/567)	5.8% (33/567)	48.5% (275/567)
	ORC-like	2.8% (8/289)	73.4% (212/289)	1.4% (4/289)	22.5% (65/289)
DCL	**SRC-like**	0% (0/4)	0% (0/4)	0% (0/4)	100% (4/4)
	ORC-like	0% (0/16)	75% (12/16)	6.3% (1/16)	18.8% (3/16)

RC structures

Table S2. Animacy of head noun and RC-internal NP in subject and object RCs for DE and DCL construction types

		RC-internal NP (animate)		RC-internal NP (inanimate)	
		Animate Head	**Inanimate Head**	**Animate Head**	**Inanimate Head**
DE	**SRC**	0% (0/79)	6.3% (5/79)	31.6% (25/79)	62.0% (49/79)
	ORC	3.7% (8/215)	74.4% (160/215)	1.4% (3/215)	20.5% (44/215)
DCL	**SRC**	0% (0/0)	0% (0/0)	0% (0/0)	0% (0/0)
	ORC	0% (0/14)	78.6% (11/14)	7.1% (1/14)	14.3% (2/14)

On the basis of the corpus data we make three crucial observations. Firstly, object RC-like and object RC structures have predominantly inanimate head nouns that contrast in animacy with the other NP, which is typically animate (about three-quarters of them: 73-78%). Thus, consistent animacy contrast cues are available to aid their interpretation in naturalistic speech. Secondly, animacy contrast cues less sharply define subject RC-like and subject RC structures; animacy configurations are more evenly distributed. Finally, in the DCL construction subject RC-like sequences are extremely rare, and subject RC structures are completely absent.

These data provide further support for the influence of children's linguistic experience on their experimental performance. All nouns in the experiment were nominally animate (i.e., animals as represented by toy referents), and thus, all things being equal, the fact that the test sentences contained animate head nouns and no contrasting animacy cues should favour subject RCs analyses *in those structures where children have evidence for the analysis*; that is, in DE RCs. However, children rarely encounter a subject DCL RC or RC-like structure that follows a subject modifying pattern. Thus, while we should expect that children may experience difficulty with both subject and object DCL RCs with no animacy cues, we should expect to see greater difficulty with subject DCL RCs. This is the pattern of results we observed.

S2. Comparing the included to the excluded children.

Figure S2 compares the proportion of correct responses and head errors by RC construction type and extraction for the included and excluded children.

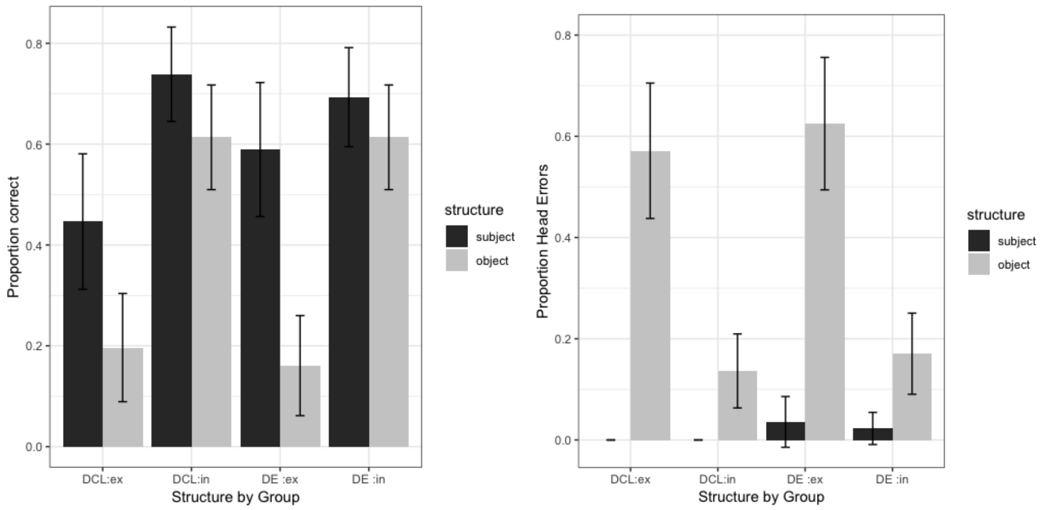

Figure S2. Proportion of correct responses (left panel) and head errors (right panel) by group (Included vs Excluded), structure (DCL versus DE) and extraction (subject versus object).

A head error occurs when children choose the RC noun as the head noun; when the errors occur in object RCs it suggests that children are pursuing a simple SVO transitive clause interpretation of the sentence. Consider sentence (1), a Mandarin DE object RC.

(1) [lao3shu3 qin1 __] de gong1ji1.

mouse kiss RL chicken.

'The chicken that the mouse kisses'.

In this sentence, a head error constitutes choosing the mouse as the referent rather than the chicken, which preserves the thematic agent-patient role assignment of the structure but fails to indicate the child understood the sentence as a head-final RC structure. Children acquiring Chinese language tend to make many head errors in object RCs (Hu, Gavarro, Vernice, & Guasti, 2016; Kidd, Chan, & Chui, 2015; Tsoi, Yang, Chan, & Kidd, 2019).

We analysed the children's performance on the task with group (i.e., included vs. excluded children) as a between-participants fixed effect using generalized linear mixed models (the *lme4* package for Linear Mixed Effects, Bates & Maechler, 2010 in R version 3.5.2; R Core Development Team, 2018). For the correct responses, the final included significant fixed effects of extraction (β = -1.63, *se*(ß) = .37, z = -4.46, p < .001), group (β = .86, *se*(ß) = .26, z = 3.33, p < .001), and a significant extraction by group interaction (β = 1.17, *se*(ß) = .4, z = 2.94, p = .003), and random intercepts for participants and items, and a random slope for group over items. The significant interaction was driven by the excluded group's large subject-object asymmetry, or in other words, their comparatively poorer performance on the object RCs.

The excluded children's poor performance on object RCs was almost entirely explained by their tendency to make head errors; that is, to choose the RC subject as the target referent (e.g., choosing the lion in *the bear that the lion pushed*). We analysed the children's head errors in the same manner as the correct responses. The final included a significant fixed effect of extraction (β = 5.57, *se*(ß) = .86, z = 6.51, p < .001) and a significant extraction by group interaction (β = -2.38, *se*(ß) = 1.13, z = -2.1, p = .036), and random intercepts for participants and items, and a random slope for group over items. In a mirror image to the correct responses, the excluded children made significantly more head errors on object RCs than the included children.

We interpret these results to indicate that the sole difference between the included and excluded children was the latter group's tendency to pursue a simple main clause analysis of object RCs. There are several reasons as to why this might be the case, such as a failure to inhibit a more frequent (SVO transitive) structure (Woodard, Pozzan, & Trueswell, 2016) or a general preference for building minimal structure (Frazier, 1987; Gibson, 2000). Since the included children's data suggest frequency-sensitive syntactic processing, the former explanation may be more likely.

S3. Analysis of offline (accuracy) responses

Children were considered to have correctly interpreted the test sentence if they selected the correct token of the head referent. These offline responses were analyzed using generalized linear mixed effects models (the *lme4* package for Linear Mixed Effects, Bates & Maechler, 2010 in R version 3.5.2; R Core Development Team, 2018). The fixed effects were: (i) sentence type (DE versus DCL), (ii) extraction (subject versus object), and (iii) their interaction. The random effects were participants and items. The maximal model had random slope for extraction under participants. Figure S3 presents children's offline correct responses to the two types of RCs (standard error bars were created by computing standard error after removing the random effects, Hohenstein & Kliegl, 2013). Overall, children were above chance at selecting the correct referent, intercept β= 0.78, z=5.7, $p < 0.001$. There were no main effects or interactions of sentence type and extraction. The accuracy of subject RCs was numerically higher than object RCs in both DE and DCL conditions, but the differences for the final included children were not significant.

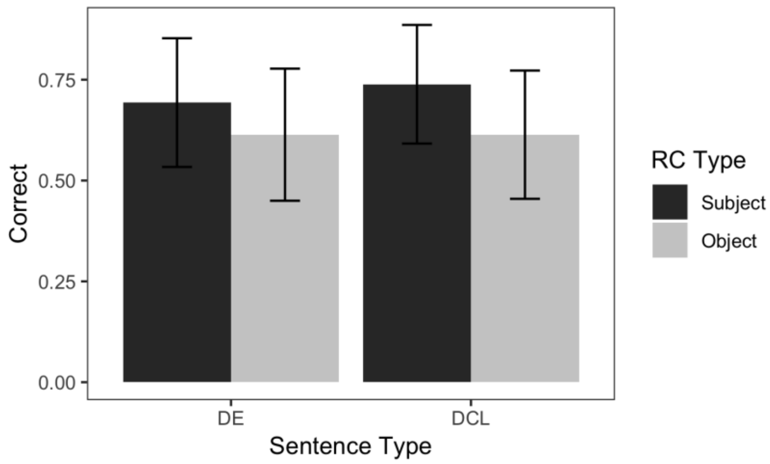

Figure S3. Accuracy of offline responses for subject and object DCL and DE RCs

S4. Comparison of permutation analyses and mixed models.

The standard approach to analysing eye-movement data is to use linear models like mixed models or ANOVA. This assumes that the DV is independent at each time sample, but typically this assumption is violated because there is often a high correlation between the behaviour at time *t* and *t*+1 (e.g., in our dataset, the correlation between target views on adjacent 40ms time windows is 0.93). One way to reduce this problem is to aggregate the data over time windows. Below we use a common approach, which is to aggregate over 200ms windows, but even with this choice we have a correlation of 0.74 between adjacent windows. Another approach would be to use autocorrelation analysis (*acf*) to find larger windows that are not correlated. However, there is still the problem that these windows are not necessarily aligned with the linguistic stimuli and may not reflect genuine parsing events, with the possibility that the same processing effect is split across two windows or that unnecessarily large windows contain genuine effects but are contaminated by unrelated noisy regions. Often researchers choose analysis windows based on prior research, but for studies like ours, for which there is no established research literature, this is not possible. To show these limitations of standard analytic approaches, we present two analyses using mixed models.

The first mixed model analysis was applied to all of the data from RC onset. First, the sum of fixations to the target referent relative to the other toys were calculated and binned into 200 millisecond windows, ranging from RC onset to 2400ms, for each participant across extraction type (subject RC, object RC) for both types of RC construction (DE, DCL). The sum of target fixations were transformed with the empirical logit.

$$elogit(p,n) = \log\left(\frac{p + 0.5}{n - p + 0.5}\right)$$

A linear mixed model was fit with a factorial combination of the fixed effects of window, extraction type, and RC construction type (all centred). The maximal model contained random intercept for subjects with no random slopes, as well as a random intercept for items with no random slopes (Barr, Levy, Scheepers, & Tily, 2013). There was a main effect of window (β=0.22, SE=0.01, $\chi^2(1)$=430.15, p<0.001) and an interaction of extraction type and RC type (β=0.47, SE=0.16, $\chi^2(1)$=8.87, p=0.0029). Crucially, the three-way window X structure X extraction interaction, the equivalent of which *was* significant in our permutation analysis, did not meet conventional significance levels (β=0.079, SE=0.04, $\chi^2(1)$=3.62, p=0.057). To illustrate further limitations of the mixed model approach, we conducted posthoc tests to understand the three-way interaction. Since every posthoc is an additional test, these tests give you an extra chance of finding a significant effect. Therefore, it is necessary to adjust for the number of comparisons, which we did automatically using the *multcomp* library (Hothorn, Bretz, & Westfall, 2008). When we did this analysis, the effect of extraction type at region 11 (z=2.35, p = 0.0188) was significant. Thus, while the permutation analysis found a significant interaction of extraction type and RC type from 2000-2400ms, the mixed model only found an effect at 2200-2400 ms. This is in part because this 400ms window was divided into two windows in the mixed model analysis and these were treated as independent effects. Finally, we point out that in this procedure we only found significant posthoc effects for the DE RC construction. This is in part because the posthocs required adjustments for multiple comparisons and this makes it more difficult to detect a significant effect.

The second mixed model analysis was the same except we began analysing the data from the first possible disambiguation point - the offset of the *de* in each sentence type. The empirical logit was equated at the offset of *de* to allow us to see how participants responded to the disambiguating input. The maximal model had a random slope for extraction type for participants and no slopes for items. There was a main effect of window, β=0.24, SE=0.03, $\chi^2(1)$=72.92, p<0.001, but no other effects were significant. This model assumes that processing behavior is purely bottom up based on the input. In contrast, the previous mixed model included all of the data from RC onset and that analysis also takes in to account the predictions and structural preferences that occur before the *de*. Since Chinese RCs are head-final, the listener accumulates a lot of important information prior to *de* that will no doubt inform parsing decisions, and this is reflected in the different outcomes of the two analyses.

In summary, the use of linear mixed models to analyse our data was less ideal than using the permutation test because (i) the mixed models analysis violates crucial assumptions of the test, (ii) analysis windows place arbitrary limits on processing effects, and (iii) multiple comparisons require *p*-value adjustments. The non-parametric permutation test overcomes both the artificial selection of time windows and problems with dependent observations. Furthermore, there are no posthoc comparisons that must be adjusted for multiple comparisons. As such, it is better suited to the analysis of time-course data yielded in eye-tracking and EEG studies (see Groppe, Urbach & Kutas, 2011 for an overview; Maris, 2012; Maris & Oostenveld, 2007; Eklund, Nichols, & Knutsson, 2016). Specifically, instead of assuming windows *a priori*, the permutation test creates clusters based on adjacent time points with significant effects and we assume that these clusters represent a processing component (e.g., comprehension of the RC which cause participants to look at the target). In the cluster, the individual time points are correlated, but they are collapsed together into one sum-*t* value for the whole cluster. The significance of the cluster is tested by randomly permutating the labels for that cluster to create an exact permutation distribution, which represents how likely it would occur by chance. See Chan et al. (2018) for a detailed tutorial about the approach.

References

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D. & Maechler, M. (2010). Matrix: sparse and dense matrix classes and methods. R package version 0.999375-39, online: <http://CRAN.R-project.org/package=Matrix>.

Chan, A., Yang, W.C., Chang, F. & Kidd, E. (2018). Four-year-old Cantonese-speaking children's online processing of relative clauses: A permutation analysis. *Journal of Child Language, 45*(1), 174-203. https://doi.org/10.1017/S0305000917000198

Chang, C. (1998). The development of autonomy in preschool Mandarin Chinese-speaking children's play narratives. *Narrative Inquiry, 8* (1), 77-111. https://doi.org/10.1075/ni.8.1.05cha

Deng, X. & Yip, V. (2018). A multimedia corpus of child Mandarin: The Tong corpus. *Journal of Chinese Linguistics, 46* (1), 69-92. 10.1353/jcl.2018.0002

Eklund, A., Nichols, T. E., & Knutsson, H. (2016). Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 201602413. https://doi.org/10.1073/pnas.1602413113

Frazier, L. (1987). Syntactic processing: evidence from Dutch. *Natural Language & Linguistic Theory*, *5*(4), 519-559.

Gibson, E. (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Y. Miyashita, A. Marantz, and W. O'Neil (Eds), *Image, language, brain* (pp.95–126). Cambridge, MA: MIT Press.

Groppe, D. M., Urbach, T. P., & Kutas, M. (2011). Mass univariate analysis of event-related brain potentials/fields I: A critical tutorial review. *Psychophysiology*, *48*(12), 1711-1725. https://doi.org/10.1111/j.1469-8986.2011.01273.x

Hohenstein, S., & Kliegl, R. (2013). *remef: Remove Partial Effects*. R package version 1.0.6.9000. https://github.com/hohenstein/remef/

Hothorn, T., Bretz, F., & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, *50*(3), 346-363. https://doi.org/10.1002/bimj.200810425

Hu, S., Gavarró, A., Vernice, M., & Guasti, M. T. (2016). The acquisition of Chinese relative clauses: contrasting two theoretical approaches. *Journal of child language*, *43*(1), 1-21. https://doi.org/10.1017/S0305000914000865

Kidd, E., Brandt, S., Lieven, E., & Tomasello, M. (2007). Object relative made easy: A crosslinguistic comparison of the constraints affecting young children's processing of relative clauses. *Language and Cognitive Processes, 22,* 860 – 897. https://doi.org/10.1080/01690960601155284

Kidd, E., Chan, A., & Chiu, J. (2015). Cross-linguistic influence in simultaneous Cantonese–English bilingual children's comprehension of relative clauses. *Bilingualism: Language and Cognition*, *18*(3), 438-452. http://dx.doi.org/10.1017/S1366728914000649

Li, X.Y. & Zhou, J. (2004). *The Effects of Pragmatic Skills of Mothers with Different Education on Children's Pragmatic Development*. Master's Thesis. Nanjing Normal University. Shanghai, China.

Mak, W. M., Vonk, W., & Schriefers, H. (2002). The influence of animacy on relative clause processing. *Journal of memory and language*, *47*(1), 50-68. https://doi.org/10.1006/jmla.2001.2837

Maris, E. (2012). Statistical testing in electrophysiological studies. *Psychophysiology, 49*(4), 549-565. https://doi.org/10.1111/j.1469-8986.2011.01320.x

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of neuroscience methods*, *164*(1), 177-190. https://doi.org/10.1016/j.jneumeth.2007.03.024

R Core Team (2018). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Online: <http://www.R-project.org>.

Tsoi, E. Y. L., Yang, W., Chan, A., & Kidd, E. (2019). Mandarin–English speaking bilingual and Mandarin speaking monolingual children's comprehension of relative clauses. *Applied Psycholinguistics, 40*(4), 933-964. https://doi.org/10.1017/S0142716419000079

Woodard, K., Pozzan, L., & Trueswell, J. C. (2016). Taking your own path: Individual differences in executive function and language processing skills in child learners. *Journal of experimental child psychology*, *141*, 187-209. https://doi.org/10.1016/j.jecp.2015.08.005

Zhou, J. (2001). *Pragmatic development of Mandarin speaking young children: from 14months to 32 months*. Unpublished doctoral dissertation, The University of Hong Kong.

Zhou, J. AcadLang Corpus. doi:10.21415/T5SC9D. Retrieved from https://childes.talkbank.org/access/Chinese/Mandarin/AcadLang.html.