Teaching English as a Second Language Quarterly (TESLQ)
(Formerly Journal of Teaching Language Skills)
44(3) 2025, pp. 55-107
https://doi.org/10.22099/tesl.2025.51984.3367
Online ISSN: 2981-1546
Research Paper

Unmasking Inconsistency in Relative Clause Ambiguity Research: A Systematic Methodological Review

Karim Vafaee Seresht [1*] **Hamideh Marefat** [2]
Abbas Ali Rezaee [3]

Abstract

Research on relative clause (RC) ambiguity resolution in first and second-language contexts has produced conflicting results, with some studies indicating a preference for high attachment, others favoring low attachment, and some reporting no clear preference. In conjunction with other variables, these mixed results may be due to variations in the methodological features employed across studies. Therefore, there is a pressing need for a systematic review of the methodological features of relevant offline tasks to evaluate how these differences may lead to conflicting results critically. To address this issue, a systematic methodological review was conducted analyzing 108 features of offline tasks, including identification, context, materials, design, administration, data analysis, open science practices, and transparency. The results revealed significant methodological variation in the literature and a moderate mean transparency score of 59.77. These findings emphasize the need for methodological standardization and greater transparency in future research to ensure reliable and comparable RC ambiguity resolution research results.

Keywords: Offline Task, Relative Clause Ambiguity Resolution, Systematic Methodological Review, Transparency Score

A growing wealth of research has been undertaken to investigate the relative clause (RC) attachment preferences of both native speakers (L1ers) and second language learners (L2ers). These RC attachment preferences have been probed employing mainly ambiguous sentences in which an RC can have, for instance, two potential host noun phrases (NPs) in the preceding complex NP construction like the ones in (1).

(1) The customer called the assistant [NP1] of the pharmacist [NP2] who was standing up.

As for L1 RC attachment preferences, some L1ers, like those of Spanish, are reported to attach ambiguous RCs (as in 1) to the first NP (NP1 or high attachment, HA; Bezerra et al., 2017; Carreiras & Clifton, 1993, 1999; Carreiras et al., 2001; García-Orza et al., 2017). Some

* Review History:
Received: 21/12/2024 Revised: 08/03/2025 Accepted: 10/03/2025

1. Ph.D., University of Tehran, Tehran, Iran; (Corresponding Author) kvafaee@ut.ac.ir
2. Professor, University of Tehran, Tehran, Iran; marefat@ut.ac.ir
3. Professor, University of Tehran, Tehran, Iran; aarezaee@ut.ac.ir

How to cite this article:
Vafaee Seresht, K. , Marefat, H. and Rezaee, A. A. (2025). Unmasking Inconsistency in Relative Clause Ambiguity Research: A Systematic Methodological Review. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *44*(3), 55-107. doi: https://doi.org/10.22099/tesl.2025.51984.3367

others, like English L1ers, attach ambiguous RCs to the second noun (NP2 or low attachment, LA; Felser et al., 2003). Yet, the results are not homogeneous as other studies are showing NP2 attachment preferences in Spanish (Carreiras et al., 2001) or when pseudo-relative structures are investigated (Alonso-Pascua, 2020; Stetie & Zunino, 2021), and no attachment preferences in some English studies (Deniz, 2022; Kim & Christianson, 2013; Tan & Foltz, 2020). As regards L2ers, more inconsistencies have been reported that should be considered. While many studies report L2ers to diverge from native-like attachment preferences (Fernández, 1999; Felser et al., 2003), some reports indicate L2ers can, in fact, display native-like preferences (Dussias, 2003; Marefat & Abdollahnejad, 2014; Marefat & Farzizadeh, 2018).

Inconsistent findings present a significant challenge, as they obscure our understanding of the underlying mechanisms responsible for the human parsing system in offline tasks. As Plonsky (2013) notes, "methodological infirmity … hinders progress in the development of theory" (p. 656). Researchers must ensure rigor and consistency in their methodological approaches to advance our understanding of the human parsing system and develop more reliable and valid theories. This underscores the need to evaluate whether the offline literature has maintained consistency in its methodological choices. If inconsistencies exist, they may be the potential moderators that have resulted in mixed results. Thus, a systematic methodological review of offline tasks provides a more comprehensive picture of the features of offline tasks, which *may* help identify potential methodological moderators. Following other syntheses (Amini Farsani & Babaii, 2020; Azadnia, 2024; Hou & Aryadoust, 2021; Liu & Brown, 2015; Marsden, Thompson et al., 2018; Plonsky, 2013, 2014; Plonsky et al., 2020; Riazi & Amini Farsani, 2024; Vafaee Seresht & Marefat, 2022), we conducted the current systematic methodological synthesis to provide a *descriptive* and *evaluative* synthesis of the abundant methodological features of offline tasks, as used in the investigation of RC attachment preferences. In this regard, Plonsky et al. (2023) indicate that "the emphasis in most methodological syntheses is on evaluating methodological practices by coding for features associated with quality, such as various design elements (e.g., sampling, random assignment), instrumentation (e.g., validity evidence, reliability), data analysis, and transparency/reporting practices" (p. 312). This review is evaluative because we focus on design features, instrumentation, data analysis, and transparency practices in the coded studies.

Methodological Reviews

Unlike substantive reviews, which aggregate the findings of primary studies to draw more definitive conclusions (Li & Wang, 2018), methodological reviews focus on the methods used to generate the findings (Marsden, Thompson et al., 2018). Essentially, methodological reviews examine the methodological features of primary research to assess whether existing practices meet specific standards and to determine potential areas for improvement (Li & Wang, 2018). Their key objectives include describing, evaluating, identifying relationships, or documenting

chronological developments and enhancements in research methodologies (Plonsky & Gonulal, 2015).

Given the calls for a 'methodological turn' (Byrnes, 2013) or 'methodological awareness' (Plonsky, 2014; Marsden, Plonsky et al., 2018) and 'methodological transparency' (Marsden, 2020) in conducting research, the number of methodological reviews has grown exponentially in applied linguistics. Many such reviews have been conducted on quantitative, qualitative, and mixed-methods research. Among them are the ones that follow.

In a notable study, Liu and Brown (2015) carried out a methodological review examining the effectiveness of corrective feedback in second language (L2) writing. They analyzed 32 published studies and 12 dissertations, focusing on the primary research's strengths and weaknesses. Their review highlighted several notable design features, such as the utilization of a 'classroom-based research' and the 'inclusive coverage of common corrective feedback strategies.' Nevertheless, they emphasized that the reviewed studies had several methodological shortcomings, including (a) insufficient descriptions of the research context, methodology, and statistical analyses; (b) experimental designs with limited generalizability; (c) the use of split-plot designs, which inhibit the identification of valid feedback effects; and (d) the use of diverse measurement tools, which make comparability of results across studies challenging.

In a methodological synthesis, Amini Farsani and Babaii (2020) retrieved and analyzed 285 unpublished MA theses in applied linguistics over 30 years. They found several shortcomings and strengths in the reviewed studies. The studies did not consistently report *p* values, used the minimum levels of confidence intervals and effect sizes, and employed low statistical power. Nevertheless, over the three decades, they also showed improvement in reporting methodological issues like reporting effect sizes and checking statistical assumptions.

In a methodological review, Morea and Ghanbar (2024) retrieved 55 empirical studies that employed Q methodology in applied linguistics. An examination of the retrieved studies' contextual, methodological, and data-analytical features indicated that the Q-sort method is gaining popularity in applied linguistics as a means to promote participant reflexivity, especially in research on teacher and learner cognition, emotions, and language-specific or multilingual motivation. Nevertheless, they also identified gaps in the employment of the Q methodology: They found frequent absence of quality-assurance measures during the creation of the Q-set and the omission of critical data-analytical details in published studies. They indicate that these shortcomings diminish the transparency and replicability of the findings derived from Q-based research.

In a methodological review of autoethnographic studies in applied linguistics, Keleş (2022) reviewed 40 autoethnographic articles published between 2010 and 2020. His review showed that many researchers used autoethnography as an umbrella term without specifying its type. Most diverged from traditional third-person academic prose yet approached their narratives analytically. However, the lack of biographical details weakened the studies' evocative and analytical depth. Furthermore, authors often failed to justify their choice of autoethnography

over other methods or explain their data collection and analysis processes. Finally, he proposes that future researchers deepen their understanding of types of autoethnography, epistemological foundations, and methodological issues to choose the more appropriate narrative approach for their reports.

Ghanbar et al. (2024) conducted a methodological systematic review of 291 narrative inquiry studies from 12 applied linguistics journals. They coded the retrieved studies based on four categories of features: (a) theoretical framework, (b) demographic features, (c) methodological features, and (d) reporting of ethics, researcher positionality, and funding status. Their review revealed a significant increase in such studies from 2012 to 2022. They also found that most studies (93%) included a theoretical framework or construct. The two most frequently reported theoretical frameworks were motivation and investment, which focus on commitment to language study and use. Most studies were conducted in the U.S. (32.1%) and focused on English (59%). Methodologically, 56% used analysis of narrative, while 32% employed narrative analysis, with thematic analysis being the dominant analytic approach (39%). However, many studies lacked transparency in reporting ethical considerations (49%), researcher positionality (55%), and funding sources (79%). The findings highlight the need for greater methodological clarity, ethical rigor, and diversity in narrative inquiry research.

Also, Plonsky and Kim (2016) systematically synthesized tasks to elicit learner language. They retrieved 85 primary studies published from 2006 to 2015. They coded the studies based on linguistic, context, and methodological features. The results revealed that the investigated language production tasks focused primarily on grammar, vocabulary, accuracy, and L2 interaction features, with limited attention to pronunciation, pragmatics, and task performance quality. A key issue was found to be a lack of theoretical and operational consistency in the field. Additionally, the data highlighted shortcomings in research and reporting practices, such as low statistical power and missing data.

Furthermore, in methodological syntheses, various methodological features are addressed. One commonly addressed aspect is what Marsden (2020) calls 'methodological transparency.' Zogmaister et al. (2024) remark that transparent reporting of the details of a scientific process is vital, as it enhances the trustworthiness of the results, reproducibility of the findings, and replicability of the research. They define methodological transparency as "clear and comprehensive documentation of the processes, techniques, and procedures employed during the study" (p. 1) such that other researchers can replicate the study without ambiguity. It is important to distinguish between 'reporting transparency' and 'methodological transparency.' Reporting transparency involves clear and comprehensive documentation of all study aspects, including the study's rationale, problem statement, research questions, hypotheses, data analysis, and methodological details (American Educational Research Association, 2006; Riazi & Amini Farsani, 2024). 'Methodological transparency' is considered a 'subset' of reporting transparency, focusing specifically on the clarity of reporting methodological issues and choices (American Educational Research Association, 2006; Øby, 2024; Wang et al., 2022).

Many systematic methodological reviews have been conducted in various domains of applied linguistics. However, no study has been conducted to investigate the methodological rigor and transparency of studies conducted on RC ambiguity resolution. To address this gap, the current methodological review has been conducted with two objectives: to identify the potential moderators that may lead to mixed results and to promote methodological rigor and transparency of future research in this domain. With such objectives in mind, we developed the following research questions.

RQ1. What specific methodological features of offline tasks have been employed in the investigation of RC ambiguity resolution?

RQ2. Have offline task features been reported transparently in the literature on RC ambiguity resolution?

Method

Study Retrieval of Offline Studies

The methodological systematic review of offline RC attachment resolution included studies from 1988 (Cuetos & Mitchell, 1988) to 2022 (Samadi et al., 2022). To retrieve as comprehensively the experimental, quantitative studies as possible and to mitigate 'publication bias (Nakanishi, 2015; Pigott, 2012), we attempted to include all 'peer-reviewed research' and 'fugitive literature' (i.e., hard-to-find literature like conference papers, M.A. theses, Ph.D. dissertations). In so doing, following Plonsky and Oswald (2015), we conducted an exhaustive keyword search in databases and search systems such as *Education Resources Information Center* (ERIC), *Linguistics and Language Behavior Abstracts* (LLBA), *Academic Search Ultimate* (ASU) *PsycINFO*, *ScienceDirect*, *Google Scholar*, *ProQuest*, *IRIS database* (Marsden et al., 2017), *Academia.edu*, and *ResearchGate.net*. Next, to retrieve any missing relevant research, we used "citation chaining" (Ziegler, 2016), "ancestry chasing" (Li & Wang, 2018), and connectedpapers.com database.

After a few trials with the keyword search, we came up with the following search terms: (offline + attachment + "relative clause"), (offline + "relative clause ambiguity"), and (offline + "relative clause resolution"). However, in *the IRIS database, these keywords yielded no* results because the studies were not indexed according to their content. Rather, since the studies were indexed based on the keywords defined by the authors, the search strategy was revised, and the search was done using words like 'processing,' 'parsing,' etc. Each search yielded a number of studies, which totaled 1,154. After removing duplicates, the results were reduced to 984 ones. Moreover, through *citation chaining* and *ancestry chasing*, the number of these studies rose to 1,007 potential studies. Upon screening the 'titles' and 'abstracts,' this number was reduced to 112. When relevance was suspected, using methodology screening, we perused the method sections which further reduced the studies to 92 ones.

Including *post-interpretive* data[2] (Kim & Christianson, 2013, 2017) from self-paced reading studies increased the number of offline studies to 99 (65 journal articles, 9 experimental book chapters, 17 conference proceedings, 3 M.A. theses, and 5 Ph.D. dissertations, see Table 1) which comprised 482 unique conditions.

Table 1

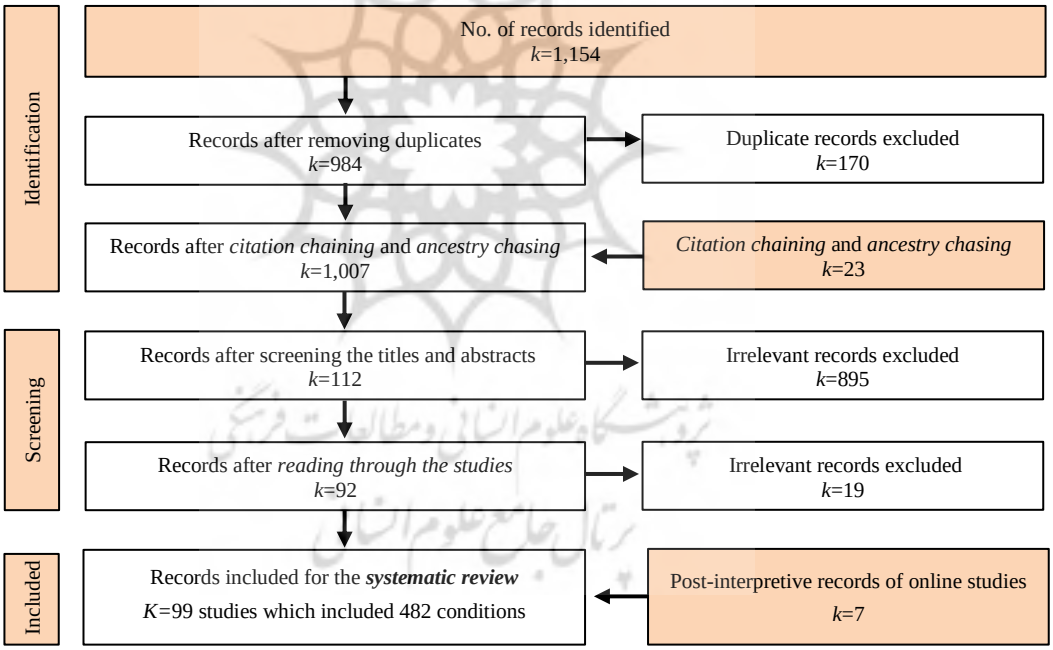
Sources of Studies and Conditions

Studies	k	%	Conditions	k	%
Journal article	66	66.67	Journal article	315	65.35
Conference paper	17	17.17	Conference paper	69	14.32
Book chapter	8	8.08	Book chapter	46	9.54
Ph.D. dissertation	5	5.05	Ph.D. dissertation	30	6.22
M.A. thesis	3	3.03	M.A. thesis	22	4.56

Note. *k* = subset of the sample; total number of studies = 99, total number of conditions = 482

Figure 1

The Retrieval Process



Inclusion and Exclusion Criteria

To minimize the 'file drawer problem' or 'publication bias' (Cooke, 2024; Rosenthal, 1979) and ensure a more comprehensive and exhaustive review, we attempted to include all relevant, accessible studies. This encompassed peer-reviewed journal articles, published works (e.g., book chapters), and fugitive literature. However, we excluded studies conducted in languages other

[2] 'Post-interpretive' results reflect cognitive processes that occur after initial parsing and interpretation of the text. For instance, comprehension questions related to RC attachment preferences is a post-interpretive elicitation technique, but in self-paced reading, reading times (RTs) reflect initial parsing or interpretation and as such are not post-interpretive.

than English as we were not proficient enough in those languages to extract the required data. Additionally, studies focusing on children and individuals with language impairments were excluded because their methodological approaches differed significantly from those of the included studies. Including them would undermine the consistency of the review.

Coding Scheme

We developed a coding manual/scheme based on the IRIS database (iris-database.org; Marsden et al., 2017) to systematically categorize the methodological features. After reviewing a random sample of 10 studies, we drafted a *coding scheme* incorporating the authors' justifications (if any) for each feature. These justifications served as a guide to resolve disagreements, such as 'small' or 'large' segmentation.

After piloting the coding scheme, we iteratively improved it, adding new features as needed. This resulted in 108 features across seven categories: (1) six identification features**,** (2) seventeen context and participant features, (3) fifty materials and design features, (4) sixteen administration and procedural features, (5) nine data analysis features, (6) four Open Science features (Marsden, Plonsky et al., 2018), and (7) six reporting transparency features (see Supplementary materials).

When the coding process finished, a second coder (one of the co-authors) coded 13.13% of the studies. Then, the intercoder reliability was calculated using Norouzian's (2021) *S* index ('meta_rate' code) for every feature (see Table 2). This enabled us to diagnose the areas of disagreement far better as the code provided us with a diagnostic report. The overall *S* index showed a sufficient agreement (*S* = 0.990).

Table 2

Sample S Index Data for the Codes

Feature	S index
Group-based vs. individual-based test item distribution	0.913
The way word concreteness is addressed	0.860
Individual- or group-based test administration	0.942
Number of lists	0.916
Type of participants	0.888
Presentation distribution of experimental and filler stimuli	0.916
Presentation instrument	0.907
Presentation type of offline tasks	0.907
Sampling type	0.797
Type of proficiency test	0.946
(Non)cumulative presentation [a]	0.907
Mean	**0.990**

Note. S indexes with total agreement (i.e., 1.00) are not reported.

[a] Non-whole item presentation: Non-whole presentation of items is a method of presenting the experimental items in a way that not the whole stimuli is presented at once, rather, they are presented word by word or chunk by chunk on the monitor.

Transparency Analysis

We employed a score-based model based on the coded features for transparency analysis. This model assesses transparency by calculating a percentage for each relevant reported feature. These individual percentages are then summed and divided by the total number of applicable features. The resulting value is multiplied by 100 to produce the mean transparency score. This method provides a clear and quantifiable measure of transparency.

Results

Results are reported based on conditions[3] rather than studies. This is justified on the grounds that (a) each condition could have been conducted in a separate study (as some studies included one condition, while some other studies included two or more conditions), (b) the results of each condition are viewed independently from other conditions in the same study, (c) there were some experiments which employed both temporarily and globally ambiguous RCs and only considering conditions independently could help us single them out for analysis purposes, and (d) basing the review on conditions provides more detail of the studies and thus the possibility of revealing potential moderators increases, which may make provide an avenue for future research.

Of the 99 reviewed studies, 9.09% were published in the journal 'Cognition', and 6.06% in 'Journal of Psycholinguistic Research' (Table 3). Also, 32.32% of the offline studies were 'fugitive literature' – conference papers, experimental book chapters, M.A. theses, or Ph.D. dissertations (labeled 'Not applicable').

Table 3

Source Journals

	k	%
Cognition	9	9.09
Journal of Psycholinguistic Research	6	6.06
Studies in Second Language Acquisition	3	3.03
Lingua	2	2.02
Quarterly Journal of Experimental Psychology	2	2.02
International Journal of Bilingualism	2	2.02
Applied Psycholinguistics	2	2.02
Other Journals[a]	41	41.41
NA for fugitive literature	32	32.32

Note. [a] Journals contributing a single study are not listed.

NA=Not applicable

[3] In (psycholinguistic) experimental studies, conditions can be defined as distinct states, levels, or values of an independent variable that are manipulated to measure their effects on a depended variable. For example, when we manipulate the 'animacy' of NPs in the complex NP to investigate its effect of on RC attachment resolution, each distinct state or level of the independent variable (i.e., animate NP1, animate NP2; animate NP1, inanimate NP2; inanimate NP1, inanimate NP2; and inanimate NP1, animate NP2) creates a distinct condition which can be compared with other conditions to investigate their (modulating) effects (see also Jegerski, 2014; Keating & Jegerski, 2015).

The distribution of these studies and conditions over time are depicted in Figures 2 and 3, respectively. In offline literature, the three years with the highest number of published *studies*, in rank order, are 2021, 2022, and 2003. As for conditions, the three highest numbers of *conditions*, in descending order, belong to 2021, 2003, and 2015.

Figure 2

Publication of Offline RC Attachment Resolution Studies over Time

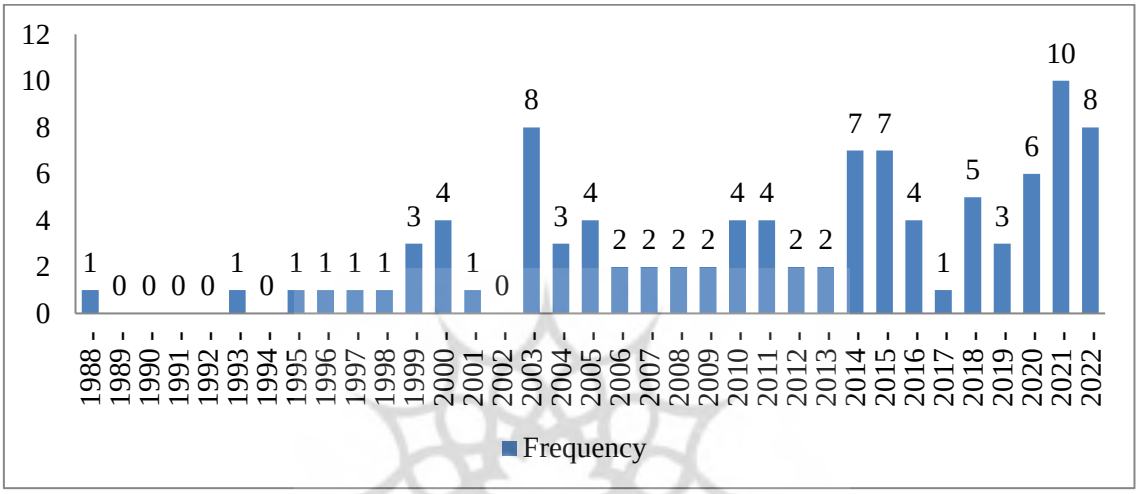
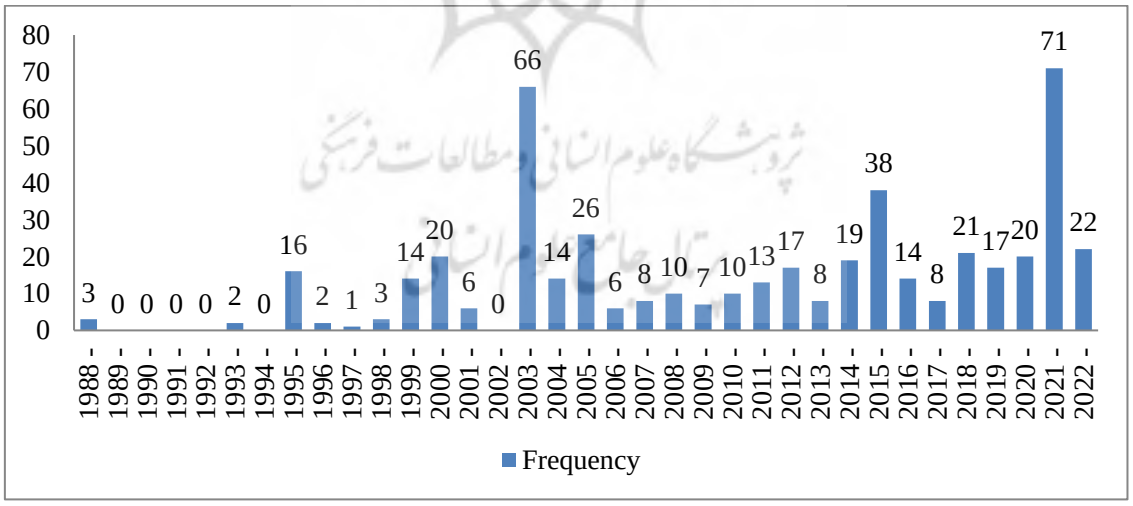


Figure 3

Publication of Offline RC Attachment Resolution Conditions over Time



Following previous systematic reviews (Plonsky & Kim, 2016; Zhang & Plonsky, 2020), we calculated feature frequencies and percentages to address the research questions.

RQ1. Offline Task Use in RC Ambiguity Resolution

RQ1 examined offline tasks as used in the literature on RC ambiguity resolution. To this end, 108 features (see Supplementary materials) were coded, including participant and context, materials and design, administration and procedural, and data analysis features.

1. Participant and Context Features

As shown in Table 4, participants come from different language backgrounds and the most common languages are English (k=100), Spanish (k=82), Korean (k=45), Turkish (k=41), and Persian (k=36).

Table 4

Number of Conditions for Different Languages

Participants' L1	k	%	Participants' L1	k	%
English	100	20.75	Hindi	4	0.83
Spanish	82	17.01	Taiwanese	4	0.83
Korean	45	9.34	Afrikaans	2	0.41
Turkish	41	8.51	Croatian	2	0.41
Persian	36	7.47	Tagalog	2	0.41
German	29	6.02	Mongolian	2	0.41
Russian	19	3.94	Spanish, Italian, German, Dutch, French,		
Spanish-English bilinguals	18	3.73	Russian, Portuguese, Greek, and Arabic	2	0.41
Italian	16	3.32	Swedish	1	0.21
French	14	2.90	Norwegian	1	0.21
Portuguese	9	1.87	Romanian	1	0.21
Greek	9	1.87	Thai	1	0.21
Japanese	9	1.87	Indonesian	1	0.21
English and Russian	8	1.66	Mandarin	1	0.21
Chinese	5	1.04	Mongolian-Chinese L3 learners of Japanese	1	0.21
Arabic	4	0.83	Kinaray-a	1	0.21
Bulgarian	4	0.83	NR	4	0.83
Dutch	4	0.83			

Note. NR = Not reported

Table 5 shows that most studies used voluntary sampling (34.23%) and convenience sampling[4] (3.11%). However, 57.47% of the conditions did not report their sampling strategy transparently. More studies can be conducted to investigate whether participant type may act as a moderator in such studies.

Moreover, most studies sampled their participants from 'university' (60.37%). However, 27.59% did not report this. Furthermore, most studies (59.34%) used 'undergraduate university students', but 19.88% did not specify participant type (Table 5). Other conditions (10.78%) drew samples from other types of participants or from a combination of university and non-

[4] A distinction is made between 'voluntary sampling' and 'convenience sampling'. In 'voluntary sampling' the participants take the initiative to participate, but in convenience sampling, the researcher takes the initiative to recruit participants.

university students. Further research may clarify whether differences in sampling context may act as a moderator in such studies.

Table 5

Some Participant and Context Features

	k	%
Sampling Type		
Voluntary	165	34.23
Convenience	15	3.11
Purposive	8	1.66
Opportunity	8	1.66
Snowball	4	0.83
Voluntary purposive	3	0.62
Homogeneous convenience sampling from volunteers	1	0.21
Stratified	1	0.21
NR	277	57.47
Institution Type		
University	291	60.37
Non-classroom	12	2.49
Various	11	2.28
University and language institute	10	2.07
School	8	1.66
School, university, and non-classroom	6	1.24
University and non-classroom	5	1.04
Language institute	4	0.83
University/school	2	0.41
NR	133	27.59
Participant Type		
Undergraduate university students	286	59.34
University students and non-university participants	15	3.11
University students and educators	12	2.49
Non-university participants	6	1.24
High school students	5	1.04
Graduates	5	1.04
School staff	4	0.83
English teachers	4	0.83
Doctoral/postdoctoral students	1	0.21
NR	144	29.88

Participants' linguistics background, a potential threat to internal validity in psycholinguistic studies, was reported only for 1.24% (6/482) of studies; 91.49% did not report this information (Table 6). To control for this internal validity threat in future research, initial screening is recommended to exclude participants with a linguistics background.

When generalizing findings to the target population, it is essential that studies have a large enough sample' (Dhivyadeepa, 2015). While sample size determination depends on several

factors (Brysbaert, 2019; Cohen et al., 2018), as a rule of thumb, a large enough sample is maintained to be 30 (Dhivyadeepa, 2015; Urdan, 2022; but see Brysbaert, 2019). This rule was taken into account in 57% (275/482) conditions (Table 6), with sample sizes ranging from 6 to 166 participants. The sample size is critical for ensuring the accuracy and statistical power of results because "the larger the sample, the greater is its chance of being representative" (Cohen et al., 2018, p. 103), thereby reducing sampling error (Ary et al., 2019). To minimize sampling error, increase power, and maximize generalizability, psycholinguistic researchers are recommended to draw large enough samples. Furthermore, variability in sample size may be a factor contributing to the inconsistency of the results in RC attachment resolution studies.

Moreover, as seen in Table 6, the mean age of participants in the retrieved studies differs significantly. Age has also been shown to be a modulator in RC attachment resolution (Frenck-Mestre & Pynte, 1997; Mardani & Modarres, 2023).

Table 6

More Participant and Context Features

	k	%
Background in Linguistics		
No	24	4.98
NA for high schoolers, and non-university participants	11	2.28
Yes, background in linguistics	6	1.24
NR	441	91.49
Sample Size (6-166)		
<30	207	42.95%
≥30	275	57.05%
Mean Age		
14-19	29	6.02
20-24	92	19.09
25-29	62	12.86
30-34	19	3.94
35-39	18	3.73
40-44	14	2.90
45-49	1	0.21
NR	247	51.24
Reporting Age range (14-49)		
Reported	153	31.74
NR	329	68.26

Table 7 shows that 64.32% conducted experiments in participants' native language settings; bilingual settings were used in only 1.24% of conditions, and only a small percentage (7.68%) did not provide such information.

Table 7 shows the participants' age of onset for L2 acquisition. In most studies, 60.58% tested participants in their L1. Additionally, early childhood was most common, but 24.69% provided no relevant information.

Table 7

Two More Participant and Context Features

	k	%
Participants' Language Learning Setting		
Native language	310	64.32
Instructed/Foreign language	75	15.56
Immersion	37	7.68
Second language	17	3.53
Bilingual	6	1.24
NR	37	7.68
Participants' Age of Onset for L2 Acquisition (Birth-16)		
From birth	2	0.41
Birth-7 years	2	0.41
Birth-8 years	2	0.41
Early childhood	16	3.32
Four	2	0.41
Ten	8	1.66
Eleven	4	0.83
Twelve	13	2.70
Thirteen	13	2.70
Fifteen	2	0.41
Twenty	1	0.21
22-33 years	2	0.41
Over Eighteen	4	0.83
NA for L1	292	60.58
NR	119	24.69

As evidenced by Karimi et al. (2021), Miyao and Omaki (2006), and Nakano (2009), L2 proficiency level can modulate their RC attachment ambiguity resolutions. Building on this evidence, the current synthesis incorporated this factor into its analysis. As shown in Table 8, the participants' proficiency levels varied from low-intermediate to advanced. Since L2 proficiency is a potential moderator, it is recommended that future studies ensure participants have comparable proficiency levels for meaningful comparisons.

Furthermore, the 'length of natural exposure to L2' can have an impact on participants' syntactic processing (Dekeyser, 2005; Dussias & Sagarra, 2007). Motivated by this line of research, participants' length of natural exposure to L2 in RC attachment resolution studies was coded and the results are shown in Table 8. L2 natural exposure ranged from below one year (1.87%) to above 10 years (6.02%). However, this feature is inapplicable for early bilinguals

(0.83%), and L1ers (k=295, 61.20%) and went unreported in 17.43% of conditions. Since L2 natural exposure is a potential moderator, further research with L2 natural exposure as a moderator or a control variable is recommended.

Table 8

L2 Proficiency and Exposure

L2 Proficiency Level	k	%	Length of Natural Exposure to L2	k	%
Intermediate	49	10.17	Below 1 year	9	1.87
Advanced	34	7.05	1-3 years	12	2.49
NA for early bilinguals	33	6.85	3-5 years	24	4.98
Upper-intermediate-advanced	14	2.9	5-7 years	12	2.49
Intermediate-advanced	9	1.87	7-10 years	6	1.24
With different proficiency levels	6	1.24	Above 10 years	29	6.02
Low-intermediate	1	0.21	No natural exposure	7	1.45
NA for L1	310	64.32	NA for early bilinguals	4	0.83
NR	26	5.39	NA for L1	295	61.20
			NR	84	17.43

Incentives shown to influence results (Chaix-Couturier et al., 2000; Weiner, 1980) included 'course credit' (8.51%), cash (7.88%), and 'both course credit and cash' (4.98%). However, 74.07% did not report incentives. For comparability purposes, using a common metric is recommended.

Table 9

Incentive for Participation

	k	%
Course credit	41	8.51
Cash	38	7.88
Course credit and cash	24	4.98
Course credit/cash	9	1.87
No compensation	9	1.87
Better grades	4	0.83
NR	357	74.07

A potential threat to the internal validity of (psycholinguistic) experiments is the confounding effect of language transfer (Kim & Christianson, 2017; Soares et al., 2022). Language transfer was controlled in 24.90% of conditions; a significant portion (67.01%) did not report this (Table 10). Researchers can mitigate this effect by excluding participants proficient in a second or a third language.

Table 10

Controlling Language Transfer Effect

	k	%
Yes	120	24.90
No	32	6.64
NA	7	1.45
NR	323	67.01

To maximize the internal validity of experiments, researchers exclude participants who might potentially distort or bias the results (for reasons listed in Table 11; see also Maroof, 2012). While this 'methodological control' was applied in 26.35% (k=127) of the conditions in offline studies, it was not applicable for 355 conditions (73.65%) as no participant was excluded.

Table 11

Participant Exclusion

	k	%
Yes, for language transfer effect and incurring more than 5 errors in responses to fillers	24	4.98
Yes, for low attention	23	4.77
Yes, for comprehension accuracy below intended threshold	23	4.77
Yes, for language transfer effect	8	1.66
Yes, for being outliers	7	1.45
Yes, for lower-than-intended comprehension accuracy and for not completing proficiency tests	6	1.24
Yes, for not completing intended tasks or not being native speakers	6	1.24
Yes, for not completing sentence completion task accurately	5	1.04
Yes, for lower-than-intended response accuracy	4	0.83
Yes, for speaking skill below 'superior'	4	0.83
Yes, for not completing intended task or for comprehension accuracy below intended threshold	3	0.62
Yes, for not doing all tasks completely	2	0.41
Yes, for being outliers or lower-than-intended response accuracy	2	0.41
Yes, for failure in WMC task	2	0.41
Yes, for language transfer effect and lower-than-intended language proficiency	2	0.41
Yes, for not completing intended tasks	2	0.41
Yes, for language disorders and bilingualism	1	0.21
Yes, for language transfer effect and lower-than-intended response accuracy	1	0.21
Yes, for scoring below 75 in grammar test	1	0.21
Yes, for incomplete answers and failure in WMC task	1	0.21
NA because nobody excluded	355	73.65

Materials and Design Features

Table 12 portrays that 74.90% of studies used researcher-developed materials; 'Adapted' and 'adopted materials' were used in 11.41% and 8.09%, respectively.

For validity purposes, even adopted materials require norming for a new population (He et al., 2021; Vannoy et al., 2011). Though 74.90% of the materials were researcher-developed, only 43.78% reported norming.

Table 12

Source and Norming of Materials

Source of Materials	k	%	Norming		
Developed	361	74.9	Yes	211	43.78
Adapted	55	11.41	No	2	0.41
Adopted	39	8.09	No, but reviewed by experts	2	0.41
Translated and adapted	18	3.73	NA	12	2.49
Translated	4	0.83	NR	255	52.90
NR	5	1.04			

The reviewed studies used primarily 'forced-choice tasks' (79.05%) followed by 'sentence completion tasks' (11.83%, Table 13). Given that task type can modulate the results (Kim & Christianson, 2013), replications using different task types are recommended to measure or control task type effect.

As for ambiguity type, global ambiguity was used in 90.87% of the conditions and 'temporary ambiguity' in 9.13% of conditions. In offline tasks, participants can regress and reread the experimental stimuli. Thus, the use of temporary ambiguity is not recommended unless the task design prevents participants from rereading the stimuli.

Table 13

Task and Ambiguity Types

	k	%
Task Type		
Forced-choice	381	79.05
Sentence completion	57	11.83
Paraphrase decision	34	7.05
Acceptability judgment	10	2.07
Ambiguity Type		
Global	438	90.87
Temporary	44	9.13

The presence and types of inter-sentential or intra-sentential prompts[5] in (non-)experimental stimuli can moderate or bias RC attachment parsing (Sokolova & Slabakova, 2019; Traxler &

[5] A prompt can be defined as a structural or non-structural cue that 'shapes RC resolution' (Sokolova & Slabakova, 2019) – i.e., causes participants to select a specific parsing strategy, especially a dispreferred one, or change their currently chosen parsing strategy. Note that 'disambiguation' of different kinds is not considered a prompt. Rather, non-disambiguation information which may somehow prompt a change in participants' RC attachment preferences is considered a prompt. For example, a semantic association between the word 'doctor' and 'examine' may prompt participants to attach the RC to NP1 in *The doctor of the patient who examined the wound died yesterday*. This

Tooley, 2007). As shown in Table 14, 24.48% of conditions contained prompts. Such prompts can be of various types: The two most prevalent types of prompts were 'syntactic' (8.30%) and 'semantic' (5.39%).

Table 14

Presence and Types of Inter- or Intra-Sentential Prompts

	k	%
Presence of Prompts		
-Prompt	364	75.52
+Prompt	118	24.48
Type of Prompts		
Syntactic priming	40	8.30
Semantic priming	26	5.39
Arousal priming	11	2.28
Structural biasing using perceptual verbs/nouns	10	2.07
Pragmatic biasing	7	1.45
Cross-domain structural priming	6	1.24
Manipulating information structure	6	1.24
Valence priming	5	1.04
Proper noun biasing	4	0.83
Implicit prosody	3	0.62
Semantic vs. morphosyntactic biasing	2	0.41
Syntactic agreement vs. pragmatic biasing	2	0.41
Implicit causality	1	0.21
NA	359	74.48

Using temporarily ambiguous stimuli, researchers need to be assured that RCs are unambiguously interpreted as referring only to one of the NPs (Jegerski, 2014; Marefat et al., 2015; Mahmoodi & Sheykholmoluki, 2022; Sokolova & Slabakova, 2021). This feature is not applicable to 88.64% of conditions using global ambiguity. Of the remaining 55 conditions, 3.32% used 'ANEW' (Affective Norms for English Words), and 8 resorted to 'norming studies' to ensure unambiguous interpretation (Table 15).

Table 15

Checking Bias in Temporarily Ambiguous Stimuli

	k	%		k	%
ANEW	16	3.32	Adding a modifier	3	0.62
Norming	8	1.66	Semantic-link.com	2	0.41
Referential co-text	4	0.83	Reflexive pronouns	2	0.41
Researcher judgment	4	0.83	Gender and number marker	1	0.21

note is 100 words; is it really needed? If not, we may say: A prompt is a cue (structural or non-structural) influencing relative clause resolution, potentially leading to the selection of a dispreferred parsing strategy, (Sokolova & Slabakova, 2019). For example, semantic association btw *doctor* and *examine* can prompt attachment preferences in *The doctor of the patient who examined the wound died yesterday*.

	k	%		k	%
Norming and expert judgment	4	0.83	Implicit causality verbs from semantic categories of 'psych' and 'judgment'	1	0.21
Using a proper noun in NP2	4	0.83	NA	427	88.59
Intra-sentential semantic biasing	4	0.83	NR	2	0.41

When employing ambiguous sentences, researchers need to ensure that attachment sites possess 'equal levels of plausibility' – both NPs in the complex NP should be 'equally natural' antecedents for RC attachment (Matić & Kovačević, 2022; Moon & Yun, 2021). Even for temporarily ambiguous sentences, both NPs should be equally plausible until the disambiguating point. As seen in Table 16, this feature was relevant in 91.08% of the conditions, yet it was addressed in 51.04% of the 439 applicable conditions. This indicates that insufficient attention is paid to 'equal levels of plausibility' in the studies.

Table 16 shows that the top three techniques to investigate equal plausibility were 'Expert judgment' (17.22%), 'Plausibility norming' (13.90%), and 'Researcher judgment' (13.07%). To investigate which of these techniques yields more plausible attachment sites, more research is required.

Table 16

Addressing Equal Plausibility of Attachment Sites

	k	%
Applicable conditions for Equal Plausibility		
Applicable for equal plausibility	439	91.08
NA for equal plausibility	43	8.92
Addressing Equal Plausibility		
Yes	246	51.04
NA for biased stimuli	40	8.3
NA for sentence completion tasks	3	0.62
NR	193	40.04
Methods for Addressing Equal Plausibility		
Expert judgment	83	17.22
Plausibility norming	67	13.9
Researcher judgment	63	13.07
Native speaker judgment of naturalness/plausibility	18	3.73
Plausibility norming and expert judgment	4	0.83
Equal plausibility NR, thus NA	191	39.63
NA for biased stimuli	40	8.3
NA for sentence completion tasks	3	0.62
Equal plausibility addressed, but not reported how	13	2.70

To develop 'truly globally ambiguous' stimuli, Başer and Hohenberger (2020) employed a novel strategy in a series of experiments. They developed and administered a set of

experimental stimuli on Turkish participants. They conducted item analysis, and those items with 'asymmetric attachment preferences' were then removed from their final experiment, which gave rise to the development of stimuli that were 'truly globally ambiguous.' This original strategy has only been conducted once in Başer and Hohenberger's (2020) study, which included two conditions (Table 17). This line of research merits further investigation.

Table 17

Removing Stimuli With Asymmetric Preferences

	k	%
No, not checked	480	99.59
Yes, stimuli with asymmetric attachment preferences removed	2	0.41

Offline RC attachment preferences can only be obtained if probe options, in the form of multiple-choice questions or fill-in-the-blanks, follow experimental stimuli. Probe questions following stimuli were used almost in all conditions (85.89% + 9.54%). A small number (0.41%) used probes for 'two-thirds of experimental stimuli' (Bidaoui et al., 2016), and 4.15% did not report use of probe options (Table 18).

Table 18

Probe Questions Use

	k	%
Yes	414	85.89
Yes, a sentence completion task	46	9.54
Two-thirds of experimental stimuli	2	0.41
NR	20	4.15

The most frequently investigated languages were English (34.65%), Spanish (19.09%), Korean (6.85%), Turkish (5.81%), and German (5.19%). As shown in Table 19, some languages are under-researched. For generalizability purposes regarding the behavior of the human parsing system, more research in these languages is required.

Table 19

Language of Experiments

	k	%		k	%		k	%
English	167	34.65	English/Russian	8	1.66	European Portuguese	2	0.41
Spanish	92	19.09	Japanese	7	1.45	Afrikaans	2	0.41
Korean	33	6.85	Arabic	5	1.04	Croatian	2	0.41
Turkish	28	5.81	Brazilian Portuguese	4	0.83	Norwegian	1	0.21
German	25	5.19	Bulgarian	4	0.83	Romanian	1	0.21
French	19	3.94	Dutch	4	0.83	Swedish	1	0.21
Russian	17	3.53	Hindi	4	0.83	Thai	1	0.21
Italian	16	3.32	Portuguese	3	0.62	Kinaray-a	1	0.21
Persian	15	3.11	Chinese	3	0.62	Tagalog	1	0.21
Greek	14	2.9	Mongolian	2	0.41			

Researchers frequently create more than one list of experimental stimuli to' maximize internal validity' of psycholinguistic experiments. This helps researchers (a) to create different conditions for 'comparability' purposes, (b) to avoid including two or more versions of the same item in one list (i.e., to avoid 'repetition effects,' Keating & Jegerski, 2015), (c) to reduce test fatigue (Marinis, 2010; Samadi et al., 2022), (d) and to avoid participants' test awareness (Samadi et al., 2022). In the reviewed studies, multiple lists of experimental stimuli were used in 57.05% of the conditions, while only one list of experimental stimuli was provided in 27.18% of the conditions (Table 20). Researchers are recommended to employ multiple lists in their psycholinguistic experiments for the above-mentioned purposes.

It is argued that "individual differences studies aim to explain as much of the variance due to individual differences as possible – while minimizing the variance due to task differences" (Swets et al., 2007, p. 67). Hence, using multiple lists is held to maximize the variance due to task differences. Following this rationale, two studies (James et al., 2018; Swets et al., 2007) report the deliberate use of 'a single list.' Other studies that employed a single list did not provide any rationale for their choice (Table 20). Researchers are called for to provide their rationales for the use of a single list or multiple lists.

Table 20
Using and Number of Lists

	k	%
Using More Than One List		
Yes	275	57.05
No	131	27.18
NR	76	15.77
Number of Lists		
One	131	27.18
Four	106	21.99
Two	103	21.37
Three	30	6.22
Six	21	4.36
Eight	12	2.49
Twelve	3	0.62
NR	76	15.77

One potential threat to the internal validity of (psycholinguistic) experiments is the 'order effect' (Brooks, 2012). To avoid such a threat, researchers use multiple versions of a test and different strategies to change the order of stimuli. In the reviewed offline studies (Table 21), this threat has been addressed using 'counterbalancing' (35.48%), 'Latin square design' (14.11%), and both 'counterbalancing and reversing the order of experimental stimuli' (0.62%). As stated, using multiple lists minimizes the 'order effect' but maximizes the variance due to task differences. Thus, researchers should decide and report the reason for their choice.

Table 21

Addressing Order Effect

	k	%
Counterbalancing	171	35.48
Latin square counterbalancing	68	14.11
Counterbalancing and reversing order of experimental stimuli	3	0.62
NA	181	37.55
NR	59	12.24

To generalize experimental findings across 'items', one should draw 'large enough samples' from the universe of items. As a rule of thumb, a 'large enough sample' is stated to be at least 30 (Dhivyadeepa, 2015; Urdan, 2022). The sample of items in the reviewed literature ranged from 3 to 74 items, with 352 (71.54%) conditions containing below-30-item samples and 120 (24.39%) conditions containing above-30-item samples (Table 22). Moreover, it is argued that with a small number of items, participants may experience 'rapid syntactic adaptation' and that "increasing the number of items may obscure adaptation effects" (Kaan & Chun, 2018, p. 97; see also Fine et al., 2013; Hopp, 2020; Malone & Mauner, 2020; Prasad & Linzen, 2021). This line of research for RC ambiguity resolution needs further research.

Table 22

Number of Experimental Stimuli in Lists

	k	%		k	%
No. of Experimental Stimuli Based on the Cut-off			**No. of Experimental Stimuli in Detail**		
<30	352	71.54	13	11	2.28
≥30	120	24.39	4	10	2.07
NR	10	2.03	14	8	1.66
No. of Experimental Stimuli in Detail			9	7	1.45
24	71	14.73	18	5	1.04
20	54	11.2	36	5	1.04
32	54	11.2	3	4	0.83
10	44	9.13	5	4	0.83
12	43	8.92	15	4	0.83
16	34	7.05	48	4	0.83
40	28	5.81	11	3	0.62
6	27	5.6	28	3	0.62
8	16	3.32	7	2	0.41
74	16	3.32	21	2	0.41
30	13	2.7	NR	10	2.07

As noted by Keating and Jegerski (2015), 'task effects' (repetition, unnatural processing) and 'participant suspicion' threaten internal validity which can be mitigated with the use of fillers. The frequency of fillers depends on the number of experimental stimuli or the participants' age (Marinis, 2010; Keating & Jegerski, 2015). Research suggests that a minimum of 50% fillers are needed (Havik et al., 2009; see Jegerski, 2014; Keating & Jegerski, 2015). As shown in Table 23, the highest ratio of fillers to experimental stimuli was 20:1, with

one experimental stimulus for every 20 fillers. The most common ratio was 2:1 (20.75%, k=100), indicating that one experimental stimulus was used with two fillers in these conditions. Additionally, the number of fillers ranged from 6 to 160 stimuli.

Table 23

Ratio and Frequency of Fillers in Lists

	k	%		k	%		k	%		k	%
Ratio of Fillers to Ex. Stimuli			**Ratio of Fillers to Ex. Stimuli**			**F. of Fillers in Each List**			**F. of Fillers in Each List**		
2.00	100	20.75	2.15	4	0.83	40	62	12.86	104	4	0.83
4.00	66	13.69	1.63	4	0.83	48	57	11.83	52	4	0.83
1.00	47	9.75	1.33	4	0.83	36	27	5.6	75	4	0.83
1.50	17	3.53	6.38	3	0.62	80	23	4.77	17	3	0.62
3.00	12	2.49	6.25	3	0.62	64	20	4.15	25	3	0.62
2.33	12	2.49	2.36	3	0.62	12	20	4.15	50	3	0.62
2.20	12	2.49	1.04	3	0.62	20	17	3.53	30	3	0.62
0.50	12	2.49	0.94	3	0.62	16	17	3.53	27	2	0.41
1.25	11	2.28	20.00	2	0.41	28	14	2.9	15	2	0.41
11.00	8	1.66	7.00	2	0.41	10	14	2.9	160	2	0.41
2.25	8	1.66	3.63	2	0.41	26	13	2.7	72	2	0.41
0.49	8	1.66	3.50	2	0.41	88	12	2.49	58	2	0.41
1.75	7	1.45	2.38	2	0.41	56	12	2.49	7	2	0.41
4.44	6	1.24	2.03	2	0.41	24	10	2.07	31	2	0.41
2.13	6	1.24	1.67	2	0.41	51	9	1.87	11	2	0.41
10.00	5	1.04	5.00	1	0.21	44	8	1.66	65	2	0.41
2.50	5	1.04	1.86	1	0.21	21	6	1.24	No filler	5	1.04
1.08	5	1.04	1.30	1	0.21	32	6	1.24	NR	59	12.24
6.00	4	0.83	1.22	1	0.21	70	6	1.24			
4.33	4	0.83	1.17	1	0.21	42	5	1.04			
3.33	4	0.83	0.79	1	0.21	60	5	1.04			
2.96	4	0.83	No filler	5	1.04	100	5	1.04			
2.92	4	0.83	NR	59	12.2	71	4	0.83			
2.34	4	0.83				6	4	0.83			

Note. F.=Frequency, Ex.=Experimental

To investigate the modulation effects of priming on participants' RC attachment resolution, about 10% of the offline studies used priming techniques; the majority of did not (90.66%). Those that did used 6 to 60 priming stimuli (Table 24). More research using priming techniques is recommended to explore the complexities of the human parsing system.

Table 24

Number of Primes

	k	%
6	20	4.15
24	14	2.90
30	2	0.41
48	5	1.04
60	4	0.83
NA	437	90.66

Task unfamiliarity can compromise the internal validity of experiments, particularly with few stimuli (Melnyk & Morrison-Beedy, 2012). To address this issue, 35.27% of the conditions incorporated practice stimuli, 0.41% did not, and only 2.90% used practice stimuli similar to the experimental stimuli (Table 25). Given the lack of sufficient attention to this issue, researchers are encouraged to incorporate practice stimuli in their experimental designs so as to minimize the effect of task unfamiliarity.

Table 25

Practice Stimuli: Presence and Similarity

Features	k	%			
Presence of Practice Stimuli			**Similarity of Practice Stimuli**		
Yes	170	35.27	Similar to fillers	27	5.60
No	2	0.41	Similar to experimental stimuli	14	2.90
NR	310	64.32	No similarity	3	0.62
			NA	283	58.71
			NR	155	32.16

A materials and design feature in offline tasks pertains to the 'sequencing and presentation' of experimental and filler stimuli. To prevent participants from deducing the study's purpose (i.e., avoiding a 'suspicion' threat to internal validity), offline tasks employed various presentation types: pseudo-randomization (31.33%), randomization (16.80%), individual randomization (10.79%), and interleaved presentation (1.45%). Notably, 39.63% of the studies did not report their presentation type. To enhance methodological transparency and facilitate replication and comparability, researchers should transparently report how stimuli are presented in their studies.

The internal validity of offline RC attachment resolution studies may be threatened by the ‘observer effect,’ where participants change their behavior simply because they are being observed (Ary et al., 2019). This effect may have a greater impact on individual-based tasks, as the researcher's presence may significantly influence participants' responses. In contrast, group-based administrations may minimize this effect. In the reviewed literature, 68.05% employed group-based task administration (Table 26). More studies in this regard are recommended to investigate whether differences in offline RC attachment resolution results from the observer effect.

Table 26

Presentation Types and Level

Features	k	%
Presentation Type of Experimental and Filler Stimuli		
Pseudo-randomization	151	31.33
Randomization	81	16.80
Individual randomization	52	10.79
Interleaved	7	1.45
NR	191	39.63
Presentation Level: Individual-Based or Fixed, Group-Based		
Fixed, group-based	328	68.05
Individual-based	81	16.80

Features	k	%
NA	11	2.28
NR	62	12.86

'Response bias' (James et al., 2018; Kane & Webster, 2013), another threat to internal validity, happens when participants favor 'Yes' or 'No' responses over the other. To mitigate this effect, researchers often counterbalance responses to questions. This strategy was employed in 37.55% of the conditions (Table 27). Researchers need to pay due attention to this internal validity threat.

One effective method for detecting and quantifying 'response bias' is the application of 'Signal Detection Theory' (Huang & Ferreira, 2020). This theory provides strategies and analytical techniques to detect and analyze the possibility of a significant 'response bias'. Of the 99 studies (482 conditions) only one study (1 condition, 0.21%, James et al., 2018) used 'signal detection' analyses (Table 27). This line of research requires further work.

Table 27

Addressing Response Bias and Signal Detection Analyses

	k	%		k	%
Addressing Response Bias			**Signal Detection Analyses**		
Yes	181	37.55	Yes	1	0.21
No	10	2.07	NA[a]	124	25.73
NA for no-response tasks	124	25.73	NR	357	74.07
NR	167	34.65			

Note. [a] Signal detection is not applicable when response bias is not addressed.

Complex NPs may modulate RC attachment resolution (De Vincenzi & Job, 1995; Gilboy & Sopena, 1996; Swets et al., 2007). Swets et al. (2007) found that English participants favor NP1 more with small (separate NP displays) than with large segmentation (both NPs together). Of the 482 conditions, 50.00% used large, while 14.11% used small segmentation. Word-by-word presentation (10.58% – not listed in the table) was categorized as small segmentation because the NPs are presented separately. The modulating effect of complex NP segmentation should also be controlled for comparability purposes.

Table 28

Complex NP Segmentation

	k	%
Large segmentation	241	50.00
Small segmentation	68	14.11
Not reported	173	35.89

'NP Gender' is another factor affecting RC attachment preferences. To measure its effect, some studies (Papadopoulou & Clahsen, 2003a) used NPs of the opposite gender (4.15%), and to control for its effect, others (Aguilar et al., 2022; Errichiello, 2021) used same-gender

complex NPs (9.96%). For some studies, this factor does not apply due to a lack of gender marking in their languages (e.g., Afrikaans, Korean, Persian, and Turkish) (Table 29).

Biosocially gendered NPs in ambiguous RCs may threaten the internal validity of results since certain professions (carpenter, secretary) are gender-specific (Ackerman, 2019; Cotter & Ferreira, 2024; Kotek et al., 2020). For instance, in the sentence '*The secretary of the princess who was killed in her office the other day ...*', 'her' may refer only to NP2 if 'secretary' is interpreted as male or to either NP1 or NP2 if interpreted as female. To safeguard internal validity, 'differential item functioning analysis' (Osterlind & Everson, 2009) or modification are required (Cotter & Ferreira, 2024). As shown in Table 29, only one study (Cotter & Ferreira, 2024) addressed this issue (0.62%). Therefore, to ensure more robust and valid results, it is recommended that future research consider this feature when constructing tasks.

Table 29

Gender and Biosocial Gender of the NPs

	k	%
Gender of Nouns in Complex NPs		
NPs of the same gender	48	9.96
NPs of opposite gender	20	4.15
NA	124	25.73
NR	290	60.17
Addressing Biosocial Gender Roles		
Yes	3	0.62
NR	479	99.38

The relation between the two nouns in the complex NP may modulate RC attachment resolution (Gilboy et al., 1995; Igoa et al., 1998). Table 30 illustrates the number of studies that included stimuli with a relationship between the nouns in the complex NP. Of the 99 studies, only three (3.03%) examined the effects of noun relations (Gilboy et al., 1995; Igoa et al., 1998; Mendelsohn & Pearlmutter, 1999). Further research is required to examine whether such a noun relationship can affect RC attachment preferences. If an effect is found, researchers should consider this feature when designing stimuli for RC attachment resolution investigations.

Table 30

Relation Between Nouns in Complex NPs

	k	%
Kinship	50	10.37
Functional	25	5.19
Functional/kinship/substance	24	4.98
Possessive	21	4.36
Functional/kinship	19	3.94
Functional/professional	18	3.73
Functional/occupational	17	3.53
Functional/kinship/occupational	14	2.9
Functional/kinship/possessive	14	2.9

	k	%
Substance	9	1.87
Occupational	8	1.66
Kinship/Possessive	4	0.83
Representational	4	0.83
Functional/kinship/occupational/possessive	4	0.83
Kinship/professional	3	0.62
Functional/substance	3	0.62
Alienable possessive	3	0.62
Part-Whole, Location-Thing, Associated With/Source From, Depiction-Depicted relation	2	0.41
Quantity/Measure	2	0.41
Inherent possession	2	0.41
Part-Whole, Location-Thing, Associated With/Source From	1	0.21
Kinship/professional, Part-Whole, Location-Thing, Associated With/Source From, Depiction-Depicted relation	1	0.21
NR	234	48.55

The Referentiality Principle (Gilboy et al., 1995) postulates ambiguous RCs attach to the (more) referential NP[6] (NPs with overt determiners like 'the'). Only two conditions (0.41%) compared referential and non-referential NPs. Also, 72.20% of conditions used complex NPs with two definite NPs, and two conditions used complex NPs with indefinite NPs (Table 31). Scant attention has been devoted to measuring and controlling the effect of referentiality.

Table 31

Referentiality

	k	%
Both NPs definite	348	72.20
Varied	8	1.66
NP1 definite, NP2 indefinite	2	0.41
Both NPs indefinite	2	0.41
NR	122	25.31

'Animacy' modulates RC attachment (Desmet & Declercq, 2006; Desmet et al., 2002, 2006; Dinçtopal-Deniz, 2010; Kwon et al., 2019). As shown in Table 32, to measure this effect, 0.41% of the conditions used 'inanimate NP1s and animate NP2s', 1.87% compared 'animate NPs' with 'inanimate NP1s and animate NP2s', 10.17% compared situations when both NPs were either animate or inanimate, and 52.28% neutralized this effect using only animate NPs and 8.51% only inanimate NPs.

Table 32

Animacy Effect

Features	k	%
Controlling Confounding Effect of Animacy		
Yes	353	73.24
No	21	4.36

[6] Referentiality can be graded. For example, NP1 may be more referential than NP2 (Hansen & Hessmann, 2015).

Features	k	%
Not reported	108	22.41
Method of Controlling Animacy Effect		
Both NPs animate	252	52.28
Both NPs animate/inanimate	49	10.17
Both NPs inanimate	41	8.51
Both NPs animate or NP1 inanimate and NP2 animate	9	1.87
Inanimate NP1 and animate NP2	2	0.41
NA	126	26.14
NR	3	0.62

To enhance internal validity, 9.54% conditions 'measured' animacy effect on RC attachment and the majority (63.69%) 'controlled' it

Table 33

Animacy as a Moderator or Control

	k	%
Animacy as a moderator variable	46	9.54
Animacy as a control variable	307	63.69
NA	129	26.76

The 46 conditions in which animacy was used as a moderator variable belong to the following studies. This shows that the animacy effect is investigated in a few languages and can be considered for more investigation. Also, if future studies confirm a significant effect, researchers should account for the moderating role of animacy when designing stimuli for RC attachment resolution tasks.

Table 34

Studies with Animacy as a Moderator Variable

Study	Investigated Language
Deniz (2022)	L1 Turkish
Başer and Hohenberger (2020)	L1 Turkish
Dinçtopal-Deniz (2010)	L1 Turkish and L2 English
Kırkıcı (2004)	L1 Turkish
Hocking (2003)	L1 English
Mitchell et al. (2000)	L1 Afrikaans
Mendelsohn and Pearlmutter (1999)	L1 English
Brysbaert and Mitchell (1996)	L1 Dutch
Gilboy et al. (1995)	L1 English and L1 Spanish
Cuetos and Mitchell (1988)	L1 English

The 'plural attraction effect', where ambiguous RCs attach to plural NPs (Aguilar et al. 2022; Lee & Garnsey, 2015; Reifegerste et al., 2020; Son, 2020), poses a potential threat to internal validity. As shown in Table 35, this effect was addressed in 0.83% of the conditions using 'either singular or plural NPs' and in 3.53% using 'singular NPs'. Future research is needed to systematically measure this effect in greater depth.

Table 35

Addressing Plural Attraction Effect

	k	%
Yes, using stimuli with both NPs singular	17	3.53
No, one NP plural and the other singular	10	2.07
Yes, using stimuli with either singular or plural NPs	4	0.83
NA for global ambiguity	437	90.66
NR	14	2.90

In a complex NP encompassing an abstract noun and a concrete noun, the concrete noun typically has a substantial memory and processing advantage. This tendency for RCs to attach to the concrete noun is known as the 'concreteness effect' (Acuña-Fariña, 2016; Ballot et al., 2022; Gardini et al., 2003; Jessen et al., 2000; Just & Brownell, 1974; Paivio, 1991). To neutralize this effect, 12.03% of the conditions deliberately employed concrete nouns and 51.04% used two animate (thus, concrete) NPs. Considering that this effect may act as a moderator in related studies, researchers should neutralize it when constructing of the stimuli. Given that this effect may act as a moderator in related studies, researchers should account for and neutralize it when designing experimental stimuli.

The 'frequency weight' of nouns in NP1 or NP2 can bias RC attachment, leading to the 'frequency effect' (García-Orza et al., 2017; Pynte & Colonna, 2001). Table 36 shows this internal validity threat was addressed in 13.07% of the conditions, though one condition (0.21%, Felser et al., 2003) did not address it, as the materials were exact translations from Papadopoulou and Clahsen (2003b). Furthermore, the frequency effect was addressed most commonly using 'Basic level words' (3.53%) and 'Davis and Perea's (2005) frequency list' (3.32%). Two key issues remain to be addressed: first, the frequency effect has not been sufficiently addressed in the literature, and second, even when addressed, no standardized metric has been consistently applied.

Table 36

Word Concreteness and Frequency Effect

Features	k	%
Addressing Word Concreteness Effect		
Both NPs animate, thus both NPs concrete	246	51.04
Both nouns concrete	58	12.03
NR	178	36.93
Addressing Word Frequency Effect in Complex NPs		
Yes	63	13.07
No	1	0.21
NR	418	86.72
Criteria for Addressing Word Frequency Effect		
Basic level words	17	3.53
Davis and Perea's (2005) frequency list	16	3.32
CELEX corpus of spoken and written English	4	0.83
Using common words	4	0.83
Using Lexique database	4	0.83

Features	k	%
A frequency ratio of at least 0.66	3	0.62
A frequency ratio of at least 0.65	2	0.41
A frequency ratio of at least 0.75	2	0.41
NA because exact translations were used	1	0.21
Davis & Gardner's (2010) frequency list	1	0.21
NR	428	88.79

The presence of 'lexical overlap' – where the lexis used in other stimuli overlaps with either of the two NPs in a complex NP – can lead to a 'priming effect' or 'attachment advantage' referred to as the 'lexical boost effect' (Kantola et al., 2023; Scheepers et al., 2017; van Gompel et al., 2022). Only three studies (Başer, 2018, 2019; Başer & Hohenberger, 2020), containing 19 conditions (3.94%), took measures to neutralize this effect (Table 37). Thus, this issue remains to be investigated and taken into account in the construction of offline tasks.

Substantial differences in the word length between NP1 and NP2 may create an 'attachment advantage' or 'sensitivity' to either NP, which can be called the 'NP length effect' (Ferreira & Clifton, 1986). This potential threat to internal validity was addressed in 7.68% of conditions. Therefore, further research is required, and greater awareness of this effect should be raised.

Table 37

Lexical Boost and NP Length Effect

	k	%
Addressing Lexical Boost Effect		
Yes	19	3.94
NR	463	96.06
Addressing NP Length Effect		
Yes	37	7.68
NR	445	92.32

Length mismatches between RCs across studies can cause conflicting results (Fernández, 2003; Hemforth et al., 2015). As shown in Table 38, many researchers have addressed RC length; however, they used varied metrics (Table 38). Thus, it is recommended that researchers use a common metric to make the results comparable.

Table 38

RC Length

	k	%
Short RC: 2 prosodic words on average	12	2.49
Long RC: 3.8 prosodic words on average	12	2.49
Short RC but metric not specified	12	2.49
Long RC but metric not specified	12	2.49
RC length between 4-6 words	6	1.24
RC length between 3-5 words	4	0.83
Long RC: 6-7 words	3	0.62
Short RC: 2.5 syllables on average	2	0.41

	k	%
Long RC: 9 syllables on average	2	0.41
Short RC: 3 syllables on average	2	0.41
Long RC: 8 syllables on average	2	0.41
Short RC: 5 syllables on average	2	0.41
Long RC: 14 syllables on average	2	0.41
Short RC: 3.5 syllables on average	2	0.41
Long RC: 11 syllables on average	2	0.41
Medium RC: 4-6 syllables	2	0.41
Short RC: 1-2 words	2	0.41
Long RC: 2-4 words	2	0.41
Long RC: Embedded verb plus a complement, and two PP adjuncts	1	0.21
Medium RC: Embedded verb plus a complement, and a PP adjunct	1	0.21
Short RC: Embedded verb plus a complement	1	0.21
RC length controlled, but not reported how	38	7.88
All types of RCs: short, medium, and long, and metric not specified	25	5.19
NR	333	69.08

The canonicality effect suggests that more frequently encountered structures (e.g., active sentences) are considered canonical, while less frequent counterparts (e.g., passive sentences) are non-canonical. Canonical structures are evidenced to be easier to process and can potentially moderate RC attachment preferences (Başer, 2018; Lim & Christianson, 2013). To address this, some researchers included have reported to use active (57.88%), passive (1.66%), or an equal number of both (5.39%) in their studies (Table 39). Due to its potential moderating effect, this effect needs to be considered in the construction of offline experimental stimuli.

Table 39

Active or Passive Stimuli

	k	%
Active	279	57.88
Active and passive	26	5.39
Passive	8	1.66
NA for sentence completion tasks	22	4.56
NR	147	30.50

Research indicates that subject-modifying RCs impose a higher processing load than object-modifying RCs (Caplan et al., 1998; Lowder & Gordon, 2021). As shown in Table 40, most conditions (71.99%) focused on object-modifying RCs, while only 27.39% examined subject-modifying RCs. To ensure comparability of results, researchers are encouraged to explore the potential moderating role of RC type. If significant, they should consider it in the construction of offline stimuli.

Table 40

Modifying RC

	k	%
Object-modifying	347	71.99
Subject-modifying	132	27.39
Both subject- and object-modifying	2	0.41
NR	1	0.21

In psycholinguistic studies, fillers serve to distract participants from the true purpose of the research (Keating & Jegerski, 2015). As displayed in Table 41, 87.34% of the conditions reported using fillers, while 1.66% (one study with 8 conditions) did not because the participants were already aware of the aim of the study (Errichiello, 2021).

To ensure L2ers' comprehension of stimuli, researchers need to assess their proficiency level. Proficiency tests were employed in 23.24% of conditions. As shown in Table 41, it was not applicable for L1ers or early bilinguals (35.89%). Table 41 also displays the different types of proficiency tests used in the offline literature.

Table 41

Fillers and Proficiency Tests

Features	k	%
Presence of Fillers		
Yes	421	87.34
No	8	1.66
Not reported	53	11.00
Presence of Proficiency Test		
Yes	112	23.24
NA	173	35.89
Not reported	197	40.87
Type of Proficiency Test		
Self-rating	33	6.85
Oxford Placement Test	13	2.70
C-test	10	2.07
Cloze Test	10	2.07
Oxford Quick Placement Test	9	1.87
Greek Language Proficiency Test used at University of Athens	6	1.24
Cloze test and TOEFL scores	4	0.83
Simulated Oral Proficiency Interview + language background self-report	4	0.83
The grammar section of the university entrance exam	4	0.83
The DELE cloze task	4	0.83
MLA reading comprehension test	3	0.62
Japanese language proficiency test	3	0.62
Grammaticality Judgment Task	2	0.41
TOEIC	2	0.41
BLP and OPT for French	2	0.41
In-house proficiency exam	1	0.21
ProTEFL	1	0.21
A battery of tests	1	0.21
NA because it includes both native and non-native participants	4	0.83
NA for early bilinguals	2	0.41
NA for L1	167	34.65
Not reported	197	40.87

The type of relativizer may modulate RC attachment (Delle Luche et al., 2006). Only one condition (conducted in French[7]) out of 482 ones investigated this effect (Table 42). This needs further research.

Table 42

Type of Relativizer

	k	%
Studies on English		
Who	75	15.56
That	39	8.09
who or that	36	7.47
that, who, which	1	0.21
Studies on Non-English Languages		
que (in Spanish)	70	14.52
ki (in Turkish a complementizer)	28	5.81
ke (in Persian, a complementizer)	15	3.11
pu (in Greek)	14	2.90
que (in Portuguese)	6	1.24
qui (in French)	5	1.04
die (in German)	4	0.83
A feminine relativizer (in Arabic)	2	0.41
die/dat (in Dutch)	2	0.41
lequel/laquelle (in French)	1	0.21
à qui (in French)	1	0.21
auquel (in French)	1	0.21
som (in Norwegian)	1	0.21
som (in Swedish)	1	0.21
care (in Romanian)	1	0.21
NR	179	37.14

Administration and Procedural Features

When offline tasks are being administered, participants may change their behavior if they feel they are being observed, a phenomenon referred to as 'the observer effect' (Ary et al., 2019). This effect could have a stronger influence on RC attachment preferences in individual-based task administrations compared to group-based ones. This is because individual-based settings may heighten participants' awareness of being observed, potentially affecting their responses. Table 43 shows that 38.17% of the conditions used individual-based and 34.02% used group-based administrations. In 3.32%, both types of administrations were used. The effect of task administration type on RC attachment is an area ripe for investigation.

[7] In some languages (Persian, Turkish), a single relativizer or complementizer prevents the investigation of this effect.

Table 43

Task Administration Type

	k	%
Individual-based	184	38.17
Group-based	164	34.02
Semi-individually	16	3.32
Individual- and group-based	16	3.32
NR	102	21.16

Poor eyesight, visual deficits, and language disorders, particularly dyslexia, may modulate RC attachment (Kristjansson & Sigurdardottir, 2022; Slaghuis et al., 1993). As Table 44 shows, 12.03% of the conditions examined vision, and 3.11% screened for language deficits. But the majority of the conditions did not provide any report in this regard. More transparent reporting is this regard is required.

Table 44

Diagnoses for Vision and Language Deficits

Features	k	%
Vision Diagnosis		
Yes	58	12.03
NR	424	87.97
Language Deficit Diagnosis		
Yes	15	3.11
NR	467	96.89

For comparability of the results across studies, details about instruments used for presenting stimuli and recording responses are absolutely needed. Table 45 illustrates that most studies (32. 57%) used ‘computer or laptop screen’ or ‘paper’ (30.91%) for presentation and ‘paper and pencil’ (32.99%) and E-prime software (8.30%) for recording responses.

Table 45

Presentation and Recording Instruments

Features	k	
Presentation Instrument		
Computer/laptop screen	157	32.57
Paper	149	30.91
A large screen	4	0.83
NR	172	35.68
Recording Instrument		
Paper and pencil	159	32.99
E-Prime	40	8.30
Software, but unspecified	33	6.85
Qualtrics	15	3.11
SuperLab	12	2.49
Linger	12	2.49
Google Forms	10	2.07
IBEX	8	1.66
Open Sesame	8	1.66

Features	k	
Presentation Instrument		
Linger/Excel spreadsheet	4	0.83
Gorilla Experiment Builder	4	0.83
MATLAB	2	0.41
TestMaker platform	2	0.41
Psyscope	2	0.41
A web-based interface	2	0.41
MATLAB, Psychophysics Toolbox, and CogToolbox	1	0.21
NR	168	34.85

Table 46 shows another procedural feature of offline tasks: 'how responses were recorded'. Participants recorded responses in 32.57% of the conditions, and in 31.95% software recorded responses.

Another procedural feature applies to tools used for recording responses. As Table 46 shows, 32.99% of conditions used 'paper and pencil', and 27.59% of conditions used 'keyboard'.

To ensure greater accuracy and comparability of the results, researchers are recommended to adopt more systematic and standardized recording strategies and tools.

Table 46

Strategies and Tools for Recording Responses

	k	%
Strategies		
Participants	157	32.57
Software	154	31.95
Google Forms	4	0.83
Web-based interfaces	2	0.41
Experimenter	2	0.41
Both software and experimenter	1	0.21
NR	162	33.61
Tools		
Paper and Pencil	159	32.99
Keyboard	133	27.59
Button Box or response pad	20	4.15
NR	170	35.27

As shown in Table 47, most stimuli (81.54%) were self-paced, and a few (3.11%) were timed. Setting individually-calibrated time limits is argued to provide more accurate results (James et al., 2018). However, none of the studies employed such time limits. This can be investigated in future studies.

Table 47 charts the different types of presentations used in the reviewed literature. Each display type has its advantages and is used for a certain function. Tan and Foltz (2020) indicate that a phrase-by-phrase display, as compared to a word-by-word presentation, facilitates reading and comprehension. Moreover, Rah (2009) supports a phrase-by-phrase display on the grounds that word-by-word displays require more concentration and higher processing capacity

on the part of readers. On the other hand, Alonso-Pascua (2020) argues that a word-by-word presentation avoids emphasizing any of the two NPs. Furthermore, whole stimulus displays are held to provide more natural processing for the participants (Logačev & Vasishth, 2016; Papadopoulou, 2006). Yet, the presentation type should fit the *study's aim*. For example, segment-by-segment presentations are required when researchers want to force a particular type of reading/processing to test a certain hypothesis, like the chunking hypothesis (Swets et al., 2007). Therefore, when the research is not aimed to test a particular hypothesis, a whole stimulus presentation is recommended because it allows for a more natural processing of stimuli (Logačev & Vasishth, 2016; Papadopoulou, 2006). As can be seen, the most commonly used presentation type is 'whole stimulus presentation' (k=308, 63.90%).

Complex NPs may split across the lines in two-line presentations, leading to prosodic phrasing and potential differences in RC attachment resolution (Clahsen & Felser, 2006; Siriwittayakorn et al., 2014; Yao & Scheepers, 2018). Thus, single-line presentations are suggested. As seen in Table 47, only 9.96% of conditions used a single-line presentation.

As for the simultaneous or non-simultaneous presentation of experimental stimuli with probe questions, there is a debate. Sokolova and Slabakova (2021) and Siriwittayakorn et al. (2015) argue against simultaneous presentation, as they think it may lead participants to notice ambiguities or reread stimuli and reconsider parsing interpretations. Conversely, Omaki (2005) supports simultaneous presentation for complex sentences or when replicating studies. Offline tasks using paper and pencil typically use simultaneous presentation, while non-simultaneous presentation is only feasible in computerized or online formats. The reviewed offline tasks almost always employ a simultaneous presentation (53.73%, Table 47).

Time-locked presentation of target stimuli, probe questions, and response options were proposed to prevent rereading and altering the initial interpretations (Cotter & Ferreira, 2024; Grillo et al., 2013; James et al., 2018). Following this line of reasoning, 16 (3.32%) conditions reported the use of time-locked presentations, often with fixed time limits (Table 47).

Table 47

Presentation Features

	k	%
Self-paced or Timed Presentation of Stimuli		
Self-paced	393	81.54
Timed	15	3.11
NR	74	15.35
Presentation Type		
Whole stimulus at once	308	63.90
Word-byword	49	10.17
Segment-by-segment	40	8.30
Region-by-region	1	0.21
NR	84	17.43
Single Line Presentation		
Yes	48	9.96
No	8	1.66

	k	%
NR	426	88.38
(Non-)Simultaneous Presentation of Stimuli and Probe Questions		
Simultaneous	259	53.73
Non-simultaneous	119	24.69
NA for sentence completion tasks	30	6.22
NR	74	15.35
Time Limit for Sentence, Question, and Response Option Presentation		
Yes	16	3.32
NA for self-paced responding	374	77.59
NR	92	19.09
Deciding Time Limit		
Piloting	8	1.66
NA for self-paced responding	374	77.59
NR	100	20.75
Individual- or Fixed, Group-Based Time Limit		
Fixed, group-based	13	2.70
NA for sentence completion, paper and pencil, and self-paced tasks	374	77.59
NR	95	19.71

Fatigue is a threat to the internal validity in lengthy tasks (Zedeck, 2014). To mitigate this effect only 1.66% (8/482) of conditions reported to have introduced obligatory or optional within-task breaks (Table 48).

Table 48

Within-Task Breaks

	k	%
Yes	8	1.66
NR	474	98.34

Research shows that 'reading modality' can moderate reading comprehension (O'Brien et al., 2014; Price et al., 2016; Robinson et al., 2019; Prior et al., 2011; Schimmel & Ness, 2017). Studies indicate that poor readers perform better with oral reading, average readers excel in silent reading, and high achievers are equally proficient in both (Miller & Smith, 1985, 1990; Schimmel & Ness, 2017). However, the reviewed studies have not taken this effect into account and have indiscriminately employed 'silent reading' in 11 (2.28%) conditions and 'reading aloud' in 97 (20.12%) conditions (Table 49).

Table 49

Reading Modality

	k	%
Silent reading	97	20.12
Reading aloud	11	2.28
NR	374	77.59

Data Analysis Features

The most commonly used techniques in reporting scores were 'percentage procedure' (82.78%) and 'ratio' procedure' (9.54%).

Table 50

Reported Scoring Procedures

	k	%
Percentage	399	82.78
Ratio	46	9.54
Mean	21	4.36
Mean z-Scored Logarithms of Raw Scores	8	1.66
Frequency	4	0.83
Regression	2	0.41
t-test	1	0.21
NR	1	0.21

Researchers try to identify and address erroneous data that lead to Type I or Type II errors (Nestor & Schutt, 2018) through different data trimming techniques to maximize internal validity. Table 51 shows that 22.20% of the conditions employed data trimming techniques and 'removed erroneous data'. The top four criteria for such removals, in descending order, included 'Stimuli with altered responses or no responses' (4.98%), 'Lack of sufficient attention' (4.36%), 'Outliers' (4.15%), and 'A threshold of comprehension accuracy' (4.15%).

As shown in Table 51, comprehension accuracy thresholds were used in 18.67% of the conditions to trim data. Among these, the 85% threshold was the most frequently used, appearing in 7.68% of the conditions.

Table 51

Data Trimming Features

Features	k	%
Data Trimming		
Yes	107	22.20
NR	375	77.80
Data Trimming Technique		
Removing	107	22.20
NA	375	77.80
Criteria for Trimming		
Stimuli with altered responses or no responses	24	4.98
Lack of sufficient attention	21	4.36
A threshold of comprehension accuracy	20	4.15
Outliers	20	4.15
Non-ambiguous classification in sentence completion tasks	8	1.66
A threshold of grammatical accuracy or incomplete task(s)	5	1.04
Incomplete task(s)	4	0.83
Incorrectly solved mathematical prime equations removed from analyses	3	0.62
Stimuli with two or no responses	2	0.41
NA	375	77.80
A Threshold Level for Comprehension Accuracy		

Features	k	%
Yes	90	18.67
No, regardless of post-stimulus comprehension accuracy	4	0.83
NR	388	80.50
What Threshold?		
85%	37	7.68
90%	19	3.94
75%	13	2.70
80%	10	2.07
95%	8	1.66
NA	392	81.33
Yes, but the threshold is NR	3	0.62

Reliability, an important feature for valid results (Ary et al., 2019; Fitzner, 2007; Ross, 2006), was reported in only 8 (8.08%) of the 99 studies (20 conditions; 4.15%), with Cronbach's α being the most frequently used (13 conditions; 2.70%; Table 52).

Table 52

Reliability Features

Features	k	%
Reliability (for studies)		
Yes	8	8.08
NR	91	91.92
Reliability (for conditions)		
Yes	20	4.15
NR	462	95.85
Reliability Index		
Cronbach's α	13	2.70
Internal consistency	3	0.62
Inter-rater reliability for sentence completion task	2	0.41
KR-20	1	0.21
Split-half reliability	1	0.21
NA	462	95.85

Despite the importance of statistical power analyses in determining sample size, as shown in Table 53, only 1.24% (k=6) of the conditions (Mahmoodi et al., 2022; Mahmoodi & Sheykholmoluki, 2022) used it.

Table 53

Statistical Power Analysis

	k	%
Yes	6	1.24
NR	476	98.76

RQ2. Methodological Transparency of Offline Tasks

Inconsistent results in studies may stem from previous studies' lack of methodological transparency (Gorgolewski & Poldrack, 2016; Marsden, 2020). Researchers often try to

replicate previous studies, but lack of methodological transparency may challenge replicability. Moreover, research validation requires replicability and confirmation of previous research (Bakken, 2019; Lindsay, 2020; Mellor et al., 2018; Miguel et al., 2014). In fact, "methodological transparency is increasingly regarded as an indicator of study quality" (Marsden, 2020, p. 26). Consequently, we investigate the degree of methodological transparency in the retrieved offline literature to evaluate the quality of research conducted in RC attachment resolution studies using a score-based model based on the coded features.

Tables 54–57 depict the extent to which the context, design, administration, and analysis features of the reviewed offline tasks have been addressed transparently.

The *mean transparency score* (TS, aka mean reporting score) for participants and context features is 59.75 (median=54.10). The small difference between the mean and the median indicating a symmetrical distribution, shows that the mean TS is a reliable indicator of central tendency.

Also, as seen in Table 54, the three features with the *lowest* TSs were 'Background in linguistics' (k=41, TS=8.51), 'Incentive for participation' (k=125, TS=25.93), and 'Age range' (k=153, TS=31.74). In contrast, exclusion criteria' (k=127, TS=100), 'Sample size' (k=481, TS=99.79), and Participants' L1 (k=478, TS=99.17) scored the highest.

Table 54

Transparency Information for Participant and Context Features

Information provided for …	k	TS
Exclusion criteria (k_{ac}=127)	127	100.00
Sample size (k_{ac}=482)	481	99.79
Participants' L1 (k_{ac}=482)	478	99.17
Participants' language learning setting (k_{ac}=482)	445	92.32
Participants' proficiency level in L2 (k_{ac}=139)	113	81.29
Type of institution (k_{ac}=482)	349	72.41
Type of participants (k_{ac}=482)	338	70.12
Participants' length of natural exposure to L2 (k_{ac}=183)	99	54.10
Mean age (k_{ac}=482)	236	48.96
Type of participation/sampling (k_{ac}=482)	205	42.53
Participants' age of onset for L2 acquisition (k_{ac}=190)	71	37.37
Avoiding language transfer effect (k_{ac}=475)	152	32.00
Age range (k_{ac}=482)	153	31.74
Incentive for participation (k_{ac}=482)	125	25.93
Background in linguistics (k_{ac}=482)	41	8.51
Mean TS		**59.75**

Note 1. k = number of reported conditions, k_{ac}= number of applicable conditions. TS = transparency score

Note 2. For features in which there are a number of 'non-applicable' conditions, after subtracting the 'non-applicable' conditions from a total number of conditions (k_t =482), the number of applicable conditions is provided in parentheses. Also, TS is obtained by dividing k by k_{ac}, and multiplying the result by 100.

Table 55 illustrates high transparency in 'design and materials' (mean TS=68.98, median=71.79); four features scored below 10 and 13 above 90.

Table 55

Transparency Information for Design and Materials Features

Information provided for …	k	TS
Task type (k_{ac}=482)	482	100.00
Ambiguity Type (k_{ac}=482)	482	100.00
Presence of prompts (k_{ac}=482)	482	100.00
Prompt type (k_{ac}=123)	123	100.00
Language of experiments (k_{ac}=482)	482	100.00
Number of primes (k_{ac}=45)	45	100.00
Modifying RC (k_{ac}=482)	481	99.79
How of controlling animacy effect (k_{ac}=356)	353	99.16
Source of materials (k_{ac}=482)	477	98.96
Number of experimental stimuli in lists (k_{ac}=482)	472	97.92
Checking bias in temporarily ambiguous stimuli (k_{ac}=55)	53	96.36
Presence of probe options for all stimuli (k_{ac}=482)	462	95.85
Addressing equal plausibility (k_{ac}=248)	235	94.76
Presence of fillers (k_{ac}=482)	429	89.00
Frequency of fillers in lists (k_{ac}=482)	423	87.76
Presentation level: individual-based or group-based (k_{ac}=471)	409	86.84
Creating more than one list (k_{ac}=482)	406	84.23
Number of lists (k_{ac}=482)	406	84.23
Addressing order effect (k_{ac}=301)	242	80.40
Controlling the confounding effect of animacy (k_{ac}=482)	374	77.59
Referentiality (k_{ac}=482)	360	74.69
Addressing plural attraction effect (k_{ac}=45)	31	68.89
Active or passive voice (k_{ac}=460)	313	68.04
Segmentation type of complex NPs (k_{ac}=482)	309	64.11
Addressing word concreteness effect (k_{ac}=482)	304	63.07
Type of relativizer (k_{ac}=482)	303	62.86
Presentation type of experimental and filler stimuli (k_{ac}=482)	291	60.37
Addressing equal plausibility (k_{ac}=432)	246	56.94
Addressing response bias (k_{ac}=358)	191	53.35
Relation type between the two nouns in complex NPs (k_{ac}=482)	248	51.45
Norming (k_{ac}=470)	215	45.74
Presence of proficiency tests (k_{ac}=309)	112	36.25
Presence of practice stimuli (k_{ac}=482)	172	35.68
RC length (k_{ac}=482)	149	30.91
Similarity of practice stimuli (k_{ac}=199)	44	22.11
Gender of nouns in complex NPs (k_{ac}=358)	68	18.99
Addressing word frequency effect of nouns in complex NPs (k_{ac}=482)	64	13.28
Criteria for addressing word frequency effect (k_{ac}=481)	53	11.02
Addressing confounding effect of word length (k_{ac}=482)	37	7.68
Addressing lexical boost effect (k_{ac}=482)	19	3.94
Addressing biosocial gender roles (k_{ac}=482)	3	0.63
Signal detection analysis (k_{ac}=358)	1	0.28
Mean TS		**68.98**

Table 56 provides transparency information for ‘administration and procedure’ (mean TS=42.58, median 43.36) with some features below 10 and some above 80.

Table 56

Transparency Information for Administration and Procedural Features

Information provided for …	k	TS
(Non-)simultaneous presentation of stimuli and probe questions (k_{ac}=452)	405	89.60
Self-paced or timed presentation of stimuli (k_{ac}=482)	408	84.65
Presentation type of tasks (k_{ac}=482)	398	82.57
Level of task administration (k_{ac}=482)	380	78.84
Strategy for recording responses (k_{ac} =482)	320	66.39
Recording instrument (k_{ac}=482)	314	65.15
Tools used for recording responses (k_{ac}=482)	312	64.73
Presentation instrument (k_{ac}=482)	310	64.32
Reading modality (k_{ac}=482)	108	22.41
Time limit for sentence, question, and response option presentation (k_{ac}=108)	16	14.82
Individual- or fixed, group-based time limit (k_{ac}=108)	13	12.04
Vision diagnosis (k_{ac}=482)	58	12.03
Single line presentation (k_{ac}=482)	56	11.62
Deciding time limits (k_{ac}=108)	8	7.40
Language deficit diagnosis (k_{ac}=482)	15	3.11
Within-task breaks (k_{ac}=482)	8	1.66
Mean TS		**42.58**

Table 57 shows high transparency in ‘data analysis’ (mean TS of 67.79). However, ‘reliability’ (mean TS of 4.14) has not received due attention.

Table 57

Transparency Information for Data Analysis Features

Information provided for …	k	TS
Data trimming technique (k_{ac}=107)	107	100.00
Criteria for trimming (k_{ac}=107)	107	100.00
Reliability index (k_{ac}=20)	20	100.00
Reported scoring procedures for RC attachment preferences (k_{ac}=482)	481	99.79
What threshold? (k_{ac}=90)	87	96.66
Data trimming (k_{ac}=482)	107	22.20
A threshold level for comprehension accuracy (k_{ac}=482)	94	19.50
Reliability (k_{ac}=482)	20	4.15
Mean TS		**67.79**

Table 58 summarizes the mean TSs of Tables 54–57, and provides an average score. Thus, based on the coded features, the reviewed offline literature enjoys a ‘total mean TS’ of 59.77.

Table 58

Total Mean TS

	TS
Design and Materials Features	68.98
Data Analysis Features	67.79
Participant and Context Features	59.75
Administration and Procedural Features	42.58
Total Mean TS	**59.77**

Discussion and Conclusion

A burgeoning growth of concern in methodological rigor, transparency, replicability, and reproducibility is witnessed in science in general (Nosek et al., 2022; Spitschan et al., 2020), and in applied linguistics, in particular (Crowther et al., 2021; Farsani et al., 2021; Hou & Aryadoust, 2021; Liu & Brown, 2015; Marsden, Thompson et al., 2018; Plonsky et al., 2020; Riazi & Amini Farsani, 2024). This rise of concern in methodological issues "is lively testimony to the fact that methodologies no longer have ancillary status in our work" (Byrnes, 2013, p. 825) and indicates that researchers are increasingly recognizing the significance of methodological issues.

With a raised awareness of the importance of methodological factors of the relevant offline literature, the current systematic review was undertaken to shed further light on two particular concerns: (a) to describe and evaluate the methodological features of the offline studies in light of the coded scheme, and (b) to investigate and evaluate the extent to which methodological issues have been reported transparently in the reviewed offline studies.

To address the first concern, the retrieved offline RC attachment ambiguity resolution literature was coded, described, and evaluated based on 108 methodological features. These features included 6 identification features, 17 context and participant features, 50 materials and design features, 16 administration and procedural features, 4 data analysis features, 4 Open Science features, and 6 transparency features. Based on the coded features, the included studies were also evaluated wherever necessary and suggestions for improvements and for future possible research were made.

To address the second concern, TSs for the reported features were calculated. As seen in Table 62, the mean TS for Open Science features was 59.77, indicating that the principles of the Open Science Framework have been implemented moderately in the reviewed offline studies. As for the most transparently reported features, there were 10 features that were reported with a perfect score of 100 (see Tables 54–57). These 10 features include 'exclusion criteria, task type, ambiguity type, presence of prompts, prompt type, the language of experiments, number of primes, data trimming technique, criteria for trimming, and reliability index'. Moreover, there were six features that were reported with a TS below five: They include the features 'reliability' (TS = 4.15), 'addressing lexical boost effect' (TS = 3.94), 'language deficit diagnosis' (TS = 3.11), 'within-task breaks' (TS = 1.66), 'addressing biosocial gender roles' (TS = 0.63), and 'signal detection analysis' (TS = 0.28). Such low TSs for these features suggest that researchers have not been attending sufficiently to the principles of Open Science.

Many attempts have been made to investigate RC attachment ambiguity resolution through offline methodology. In this regard, studies have tried to employ similar or the same materials and methodology as those used by previous ones. However, as stated previously, such attempts have *partially* failed to replicate the same results. Researchers have attributed these partial replication failures to cross-linguistic variability, individual differences, or methodological variations from non-transparent reporting practices or lacking methodological standards

(Boegle et al., 2021; Marsden, 2020). The current systematic review aimed to depict and evaluate the existing methodological variations in the offline literature on RC attachment ambiguity resolution. As shown, great methodological variations (e.g., mean age of participants, participants' age of onset for l2 acquisition, length of natural exposure to L2, methods for addressing equal plausibility, number of experimental stimuli, ratio of fillers, RC length) existed in these studies, which might explain some portion of the variability of the results. Furthermore, some features (e.g., reliability) were under-addressed.

Limitations of the Review

There are some limitations associated with the methodological review we carried out. First, regarding the scope of the review, the review focused exclusively on offline tasks, which measure final interpretations of ambiguous RCs but lack real-time processing insights captured by online methods like eye-tracking. It also excluded studies involving children and individuals with language impairments, limiting generalizability to these populations. Future reviews should expand to include online tasks and diverse participant groups. Second, the review was limited to studies published in English, which may have resulted in the exclusion of relevant research conducted in other languages. To address this limitation, future reviews should involve researchers proficient in other languages to ensure a more comprehensive inclusion of experimental studies from non-English sources. This approach would enhance the breadth and representativeness of the systematic review. Third, transparency scores were calculated based on the presence or absence of reported methodological features. However, the depth of the reporting was not assessed. For example, a study might briefly mention a methodological feature without providing sufficient detail, yet it would still be counted as "reported." This approach may overestimate the actual transparency of some studies. Finally, as this review was part of a larger project, we had to establish an ending point and finish the study retrieval process. Future systematic reviews could include additional studies to expand the scope. The review encompassed studies published between 1988 and 2022, providing a broad timeframe for comprehensive analysis. However, methodological practices and reporting standards have evolved over time, meaning older studies may have lower transparency scores due to outdated reporting practices. To address this, the timeframe could be divided into smaller segments (e.g., three periods) to compare methodological transparency across different eras, offering deeper insights into how methodological transparency practices have evolved over time.

Implications and Suggestions for Future Research

Our methodological review carries several important implications. First, the review revealed substantial methodological heterogeneity across studies. This variation highlights the need for standardized protocols in RC ambiguity research to enhance comparability across studies. Second, the review identified several methodological features that may act as moderators in RC ambiguity resolution, including task type, ambiguity type, and the presence

of prompts. For example, the use of forced-choice tasks was predominant, but the impact of task type on attachment preferences remains underexplored. Similarly, syntactic or semantic prompts were found to influence RC attachment, yet only a small percentage of studies explicitly addressed this issue. Future research should systematically investigate these and other moderators to better understand their impact on RC attachment preferences. Third, many studies failed to adequately control for confounding variables such as animacy, word frequency, and referentiality, which are known to influence RC attachment. For instance, while some studies controlled for animacy by using only animate nouns, others did not report how they addressed this potential confound. Similarly, the frequency effect of nouns in complex NPs was rarely addressed, and when it was, no standardized metric was consistently applied. Future studies should adopt more rigorous controls for these variables to ensure the internal validity of their findings. Finally, the moderate mean transparency score (59.77) indicates that many studies lack sufficient detail in reporting key methodological features. For example, critical information such as participants' linguistic background, incentives for participation, age range, and reliability were often underreported. This lack of transparency hinders the replicability of studies and limits the ability of researchers to identify potential moderators of RC attachment preferences. Future studies should prioritize transparent reporting, particularly in areas such as participant demographics, task design, and task administration.

Acknowledgments

We would like to thank the editorial team of TESL Quarterly for granting us the opportunity to submit the current systematic methodological review. We would also like to express our appreciation to the anonymous reviewers for their careful, detailed reading of our manuscript and their many insightful comments and suggestions.

Declaration of conflicting interests

The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for this article's research, authorship, and/or publication.

Supplementary Materials

The link to the supplementary materials is as follows: https://osf.io/mdu9n

References

Ackerman, L. M. (2019). Syntactic and cognitive issues in investigating gendered coreference. *Glossa: A Journal of General Linguistics*, *4*(1),117. https://doi.org/10.5334/gjgl.721

Acuña-Fariña, J. C. (2016). (A few) psycholinguistic properties of the NP. In K. Davidse (Ed.), *The structure of the English NP: Synchronic and diachronic explorations. Functions of language*, (pp. 120-141). John Benjamins. https://doi.org/10.1075/fol.23.1.01acu

Aguilar, M., Ferré, P., Hinojosa, J. A., Gavilán, J. M., & Demestre, J. (2022). Locality and attachment preferences in preverbal versus post-verbal Relative Clauses. *Language, Cognition and Neuroscience*, *37*(10), 1303-1310. https://doi.org/10.1080/23273798.2022.2066701

Alonso-Pascua, B. (2020). New evidence on the pseudorelative-first hypothesis: Spanish attachment preferences revisited. *Topics in Linguistics*, *21*(1), 15-44. https://doi.org/10.2478/topling-2020-0002

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher*, *35*(6), 33-40. https://doi.org/10.3102/0013189X035006033

Amini Farsani, M., & Babaii, E. (2020). Applied linguistics research in three decades: A methodological synthesis of graduate theses in an EFL context. *Quality & Quantity*, *54*(4), 1257-1283. https://doi.org/10.1007/s11135-020-00984-w

Ary, D., Jacobs, L. C., Irvine, C. K., & Walker, D. A. (2019). *Introduction to research in education* (10th ed.). Cengage Learning.

Azadnia, M. (2024). ChatGPT-assisted language learning and teaching: A Scoping review of research on ChatGPT use in L2 pedagogy and education. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *43*(2), 49-85. https://doi.org/10.22099/tesl.2024.49169.3250

Bakken, S. (2019). The journey to transparency, reproducibility, and replicability. *Journal of the American Medical Informatics Association*, *26*(3), 185-187. http://doi.org/10.1093/jamia/ocz007

Ballot, C., Robert, C., & Mathey, S. (2022). Word imageability influences the emotionality effect in episodic memory. *Cognitive Processing*, *23*(4), 655-660. https://doi.org/10.1007/s10339-022-01102-4

Başer, Z. (2018). *Syntactic priming of relative clause attachment in monolingual Turkish speakers and Turkish learners of English* [Unpublished doctoral dissertation]. Middle East Technical University. https://open.metu.edu.tr/bitstream/handle/11511/27252/index.pdf

Başer, Z. (2019). A universal parser or language specific parsing strategies: A study on relative clause attachment preference in Turkish. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, (16), 1-21. https://doi.org/10.29000/rumelide.648403

Başer, Z., & Hohenberger, A. (2020). Is there a particular RC attachment preference in Turkish? Negotiating the effects of semantic factors. *Journal of Psycholinguistic Research*, *49*(4), 511–539. https://doi.org/10.1007/s10936-020-09698-4

Bezerra, G. B., Leitão, M. M., & Medeiros, L. D. S. N. (2017). The influence of referentiality on relative clause processing in Brazilian Portuguese. *Revista de Estudos da Linguagem*, *25*(3), 1397-1431. http://oaji.net/articles/2019/3404-1557494834.pdf

Bidaoui, A., Foote, R., & Abunasser, M. (2016). Relative clause attachment in native and L2 Arabic. *International Journal of Arabic Linguistics*, *2*(2), 75-95.

Boegle, R., Gerb, J., Kierig, E., Becker-Bense, S., Ertl-Wagner, B., Dieterich, M., & Kirsch, V. (2021). Intravenous delayed gadolinium-enhanced MR imaging of the endolymphatic space: a methodological comparative study. *Frontiers in Neurology*, *12*, 360. https://doi.org/10.3389/fneur.2021.647296

Brooks, J. L. (2012). Counterbalancing for serial order carryover effects in experimental condition orders. *Psychological Methods*, *17*(4), 600–614. https://doi.org/10.1037/a0029310

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. Journal of Cognition, 2(1), 16. http://doi.org/10.5334/joc.72

Byrnes, H. (2013). Notes from the editor. *Modern Language Journal*, 97, 825–827. https://doi.org/10.1111/j.1540-4781.2013.12051.x

Caplan, D., Alpert, N., & Waters, G. (1998). Effects of syntactic structure and propositional number on patterns of regional cerebral blood flow. *Journal of Cognitive Neuroscience*, *10*(4), 541-552. https://doi.org/10.1162/089892998562843

Carreiras, M., & Clifton, C. (1993). Relative clause interpretation preferences in Spanish and English. *Language and Speech*, *36*(4), 353-372. https://doi.org/10.1177/002383099303600401

Carreiras, M., & Clifton, C. (1999). Another word on parsing relative clauses: Eyetracking evidence from Spanish and English. *Memory & Cognition*, *27*(5), 826-833. https://doi.org/10.3758/bf03198535

Carreiras, M., Betancort, M., & Meseguer, E. (2001, March). *Relative clause attachment in Spanish: Do readers use different strategies when disambiguating by gender and number* [Poster presentation]. Presented at the 14th Annual CUNY Conference on Human Sentence Processing. University of Pennsylvania, Philadelphia, PA.

Chaix-Couturier, C., Durand-Zaleski, I., Jolly, D., & Durieux, P. (2000). Effects of financial incentives on medical practice: results from a systematic review of the literature and methodological issues. *International Journal for Quality in Health Care*, *12*(2), 133-142.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3-42. https://doi.org/10.1017/S0142716406060024

Cohen, L., Manion, L., & Morrison, K. (2018). *Research methods in education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Cooke, R. (2024). *Meta-Analysis for Psychologists*. Springer Nature Switzerland.

Cotter, B.T., Ferreira, F. (2024). The relationship between working memory capacity, bilingualism, and ambiguous relative clause attachment. *Memory and Cognition*, 52, 1530–1547. https://doi.org/10.3758/s13421-024-01561-4

Crowther, D., Kim, S., Lee, J., Lim, J., & Loewen, S. (2021). Methodological synthesis of cluster analysis in second language research. *Language Learning*, *71*(1), 99-130. https://doi.org/10.1111/lang.12428

Cuetos, F., & Mitchell, D. C. (1988). Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in Spanish. *Cognition*, *30*(1), 73-105. https://doi.org/10.1016/0010-0277(88)90004-2

De Vincenzi, M., & Job, R. (1995). An investigation of late closure: The role of syntax, thematic structure, and pragmatics in initial interpretation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1303. https://doi.org/10.1037/0278-7393.21.5.1303

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? A review of issues. *Language learning*, *55*. https://doi.org/10.1111/j.0023-8333.2005.00294.x

Delle Luche, C., van Gompel, R. P., Gayraud, F., & Martinie, B. (2006, August). *Effect of relative pronoun type on relative clause attachment* [Paper presentation]. Presented at the Ambiguity in Anaphora Workshop Proceedings, Málaga, Spain.

Deniz, N. D. (2022). Processing syntactic and semantic information in the L2: Evidence for differential cue-weighting in the L1 and L2. *Bilingualism: Language and Cognition*, *25*(5), 713-725. https://doi.org/10.1017/S1366728921001140

Desmet, T., & Declercq, M. (2006). Cross-linguistic priming of syntactic hierarchical configuration information. *Journal of Memory and Language*, *54*(4), 610–632. https://doi.org/10.1016/j.jml.2005.12.007

Desmet, T., De Baecke, C., & Brysbaert, M. (2002). The influence of referential discourse context on modifier attachment in Dutch. *Memory & Cognition*, *30*(1), 150-157. https://doi.org/10.3758/BF03195274

Desmet, T., De Baecke, C., Drieghe, D., Brysbaert, M., & Vonk, W. (2006). Relative clause attachment in Dutch: On-line comprehension corresponds to corpus frequencies when lexical variables are taken into account. *Language and Cognitive Processes*, *21*(4), 453-485. https://doi.org/10.1080/01690960400023485

Dhivyadeepa, E. (2015). *Sampling techniques in educational research*. Lulu.

Dinçtopal-Deniz, N. (2010). Relative clause attachment preferences of Turkish L2 speakers of English. In B. VanPatten, & J. Jegerski (Eds.), *Research in second language processing and parsing*, (pp. 27-63). John Benjamins. https://doi.org/10.1075/lald.53.02din

Dussias, P. E. (2003). Syntactic ambiguity resolution in L2 learners: Some effects of bilinguality on L1 and L2 processing strategies. *Studies in Second Language Acquisition*, *25*(4), 529-557. https://doi.org/10.1017/S0272263103000238

Dussias, P. E., & Sagarra, N. (2007). The effect of exposure on syntactic parsing in Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *10*(1), 101-116. https://doi.org/10.1017/S1366728906002847

Errichiello, D. (2021). *Shall I take the high node? Cross-linguistic structural priming of relative clause attachment in Italian-English late bilinguals* [Master's thesis]. Università Ca' Foscari Venezia.

Farsani, M. A., Jamali, H. R., Beikmohammadi, M., Ghorbani, B. D., & Soleimani, L. (2021). Methodological orientations, academic citations, and scientific collaboration in applied linguistics: What do research synthesis and bibliometrics indicate?. *System*, *100*, 102547. https://doi.org/10.1016/j.system.2021.102547

Felser, C., Roberts, L., Marinis, T., & Gross, R. (2003). The processing of ambiguous sentences by first and second language learners of English. *Applied Psycholinguistics*, *24*(3), 453-489. https://doi.org/10.1017/S0142716403000237

Fernández, E. M. (1999). Processing strategies in second language acquisition: Some preliminary results. In E. C. Klein & G. Martohardjono (Eds.), *The development of second language grammars: A generative approach* (pp. 217-239). Amsterdam, NL: John Benjamins. https://doi.org/10.1075/lald.18.12fer

Fernández, E. M. (2003). *Bilingual sentence processing: Relative clause attachment in English and Spanish*. John Benjamins. http://doi.org/10.1017/S0272263104320048

Ferreira, F., & Clifton Jr, C. (1986). The independence of syntactic processing. *Journal of Memory and Language*, *25*(3), 348-368. https://doi.org/10.1016/0749-596X(86)90006-9

Fine, A. B., Jaeger, T. F., Farmer, T. A., & Qian, T. (2013). Rapid expectation adaptation during syntactic comprehension. *PloS One*, *8*(10), e77661. https://doi.org/10.1371/journal.pone.0077661

Fitzner, K. (2007). Reliability and validity a quick review. *The Diabetes Educator*, *33*(5), 775-780. https://doi.org/10.1177/0145721707308172

Frenck-Mestre, C., & Pynte, J. (1997). Syntactic ambiguity resolution while reading in second and native languages. *The Quarterly Journal of Experimental Psychology A*, *50*(1), 119-148.

García-Orza, J., Gavilán, J. M., Fraga, I., & Ferré, P. (2017). Testing the online reading effects of emotionality on relative clause attachment. *Cognitive Processing*, *18*(4), 543-553. https://doi.org/10.1007/s10339-017-0811-z

Gardini, S., De Beni, R., & Cornoldi, C. (2003). Can we have an image of a concept? The generation process of general and specific mental images. *Imagination, Cognition and Personality*, *23*(2), 193–200. https://doi.org/10.2190/j7qc-t5cg-5fkw-xcmx

Ghanbar, H., Cinaglia, C., Randez, R. A., & De Costa, P. I. (2024). A methodological synthesis of narrative inquiry research in applied linguistics: What's the story?. *International Journal of Applied Linguistics*, *34*(4), 1629-1655. https://doi.org/10.1111/ijal.12591

Gilboy, E., & Sopena, J. M. (1996). Segmentation effects in the processing of complex noun pronouns with relative clauses. In M. Carreiras, J. E. Garcia-Albea, & N. Sebastian (Eds.), *Language processing in Spanish* (pp. 191-206). Hillsdale, NJ: Erlbaum.

Gilboy, E., Sopena, J. M., Clifton Jr., C., & Frazier, L. (1995). Argument structure and association preferences in Spanish and English complex NPs. *Cognition*, *54*(2), 131-167.

Gorgolewski, K. J., & Poldrack, R. A. (2016). A practical guide for improving transparency and reproducibility in neuroimaging research. *PLoS Biology*, *14*(7), e1002506.

Grillo, A., Tomaz, M., Gomes, M. D. C. L., & Santi, A. (2013). *Pseudo Relatives vs. Relative Clauses* [Poster presentation]. Presented at the 19th Annual Conference on Architectures and Mechanisms for Language Processing (AMLaP), Marseille, France.

Hansen, M., & Hessmann, J. (2015). Researching linguistic features of text genres in a DGS corpus: The case of finger loci. *Sign Language & Linguistics*, *18*(1), 1-40. http://doi.org/10.1075/sll.18.1.01han

Havik, E., Roberts, L., Van Hout, R., Schreuder, R., & Haverkort, M. (2009). Processing subject-object ambiguities in the L2: A self-paced reading study with German L2 learners of Dutch. *Language Learning*, *59*(1), 73-112. https://doi.org/10.1111/j.1467-9922.2009.00501.x

He, X., Chen, P., Wu, J., & Dong, Z. (2021). Deep learning-based teaching strategies of ideological and political courses under the background of educational psychology. *Frontiers in Psychology*, *12*, 731166. https://doi.org/10.3389/fpsyg.2021.731166

Hemforth, B., Fernández, S., Clifton Jr., C., Frazier, L., Konieczny, L., & Walter, M. (2015). Relative clause attachment in German, English, Spanish and French: Effects of position and length. *Lingua*, *166*, 43-64. https://doi.org/10.1016/j.lingua.2015.08.010

Hopp, H. (2020). Morphosyntactic adaptation in adult L2 processing: Exposure and the processing of case and tense violations. *Applied Psycholinguistics*, *41*(3), 627-656. https://doi.org/10.1017/S0142716420000119

Hou, Z., & Aryadoust, V. (2021). A review of the methodological quality of quantitative mobile-assisted language learning research. *System*, *100*, 102568. https://doi.org/10.1016/j.system.2021.102568

Huang, Y., & Ferreira, F. (2020). The application of signal detection theory to acceptability judgments. *Frontiers in Psychology*, *11*, 73. https://doi.org/10.3389/fpsyg.2020.00073

Igoa, J. M., Carreiras, M., & Meseguer, E. (1998). A study on late closure in Spanish: Principle-grounded vs. frequency-based accounts of attachment preferences. *The Quarterly Journal of Experimental Psychology Section A*, *51*(3), 561-592. https://doi.org/10.1080/713755775

James, A. N., Fraundorf, S. H., Lee, E. K., & Watson, D. G. (2018). Individual differences in syntactic processing: Is there evidence for reader-text interactions?. Journal of Memory and Language, 102, 155-181. https://doi.org/10.1016/j.jml.2018.05.006

Jegerski, J. (2014). Self-Paced Reading. In B. VanPatten & J. Jegerski (Eds.), *Research methods in second language psycholinguistics* (pp. 20–49). Routledge.

Jessen, F., Heun, R., Erb, M., Granath, D. O., Klose, U., Papassotiropoulos, A., & Grodd, W. (2000). The concreteness effect: Evidence for dual coding and context availability. *Brain and Language*, *74*(1), 103-112. https://doi.org/10.1006/brln.2000.2340

Just, M. A., & Brownell, H. H. (1974). Retrieval of concrete and abstract prose descriptions from memory. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *28*(3), 339-350. https://doi.org/10.1037/h0082000

Kaan, E., & Chun, E. (2018). Syntactic adaptation. In K. D. Federmeier & D. G. Watson (Eds.), *Psychology of Learning and Motivation* (Vol. 68, pp. 85-116). Academic Press. https://doi.org/10.1016/bs.plm.2018.08.003

Kane, J. E., & Webster, G. D. (2013). Heuristics and biases that help and hinder scientists: toward a psychology of scientific judgment and decision making. In G. J. Feist, & M. E. Gorman (Eds.), *Handbook of the psychology of science*, (pp. 437-459). Springer Publishing Company.

Kantola, L., van Gompel, R. P., & Wakeford, L. J. (2023). The head or the verb: Is the lexical boost restricted to the head verb?. *Journal of Memory and Language*, *129*, 104388. https://doi.org/10.1016/j.jml.2022.104388

Karimi, M. N., Samadi, E., & Babaii, E. (2021). Relative Clause Attachment Ambiguity Resolution in L1-Persian Learners of L2 English: The Effects of Semantic Priming and Proficiency. *Journal of Modern Research in English Language Studies*, *8*(3), 153-185. http://doi.org/10.30479/jmrels.2020.13469.1666

Keating, G. D., & Jegerski, J. (2015). Experimental designs in sentence processing research: A methodological review and user's guide. *Studies in Second Language Acquisition*, *37*(1), 1-32. https://doi.org/10.1017/S0272263114000187

Keleş, U. (2022). Autoethnography as a recent methodology in applied linguistics: A methodological review. *Qualitative Report*, *27*(2). https://doi.org/10.46743/2160-3715/2022.5131

Kim, J. H., & Christianson, K. (2013). Sentence complexity and working memory effects in ambiguity resolution. *Journal of Psycholinguistic Research*, *42*(5), 393-411. https://doi.org/10.1007/s10936-012-9224-4

Kim, J. H., & Christianson, K. (2017). Working memory effects on L1 and L2 processing of ambiguous relative clauses by Korean L2 learners of English. *Second Language Research*, *33*(3), 365-388. https://doi.org/10.1177/0267658315623322

Kotek, H., Babinski, S., Dockum, R., & Geissler, C. (2020). Gender representation in linguistic example sentences. *Proceedings of the Linguistic Society of America*, *5*(1), 514-528. http://doi.org/10.3765/plsa.v5i1.4723

Kristjansson, A., & Sigurdardottir, H. M. (2022). *The role of visual factors in dyslexia*. PsyArXiv. https://doi.org/10.31234/osf.io/n8xer

Kwon, N., Ong, D., Chen, H., & Zhang, A. (2019). The role of animacy and structural information in relative clause attachment: evidence from Chinese. *Frontiers in Psychology*, *10*, 1576. https://doi.org/10.3389/fpsyg.2019.01576

Lee, E. K., & Garnsey, S. M. (2015). An ERP study of plural attraction in attachment ambiguity resolution: Evidence for retrieval interference. *Journal of Neurolinguistics*, *36*, 1-16. https://doi.org/10.1016/j.jneuroling.2015.04.004

Li, S., & Wang, H. (2018). Traditional literature review and research synthesis. In A. Phakiti, P. De Costa, L. Plonsky, & S. Starfield (Eds.), *The Palgrave handbook of applied linguistics research methodology* (pp. 123-144). Springer.

Lim, J. H., & Christianson, K. (2013). Second language sentence processing in reading for comprehension and translation. *Bilingualism: Language and Cognition*, *16*(3), 518-537. https://doi.org/10.1017/S1366728912000351

Lindsay, D. S. (2020). Seven steps toward transparency and replicability in psychological science. *Canadian Psychology/Psychologie Canadienne*, *61*(4), 310-317. https://doi.org/10.1037/cap0000222

Liu, Q., & Brown, D. (2015). Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing*, *30*, 66-81. https://doi.org/10.1016/j.jslw.2015.08.011

Logačev, P., & Vasishth, S. (2016). A multiple-channel model of task-dependent ambiguity resolution in sentence comprehension. *Cognitive Science*, *40*(2), 266-298. https://doi.org/10.1111/cogs.12228

Lowder, M. W., & Gordon, P. C. (2021). Relative clause effects at the matrix verb depend on type of intervening material. *Cognitive Science*, *45*(9), e13039. https://doi.org/10.1111/cogs.13039

Mahmoodi, M. H., & Sheykholmoluki, H. (2022). Working Memory Capacity and Semantic-Morphosyntactic Competition: A Comparison of L1 and L2 Sentence Processing. *Two Quarterly Journal of English Language Teaching and Learning University of Tabriz*, *14*(29), 79-98. http://doi.org/10.22034/elt.2022.50559.2481

Mahmoodi, M. H., Sheykholmoluki, H., Zoghipaydar, M. R., & Shahsavari, S. (2022). Working memory capacity and relative clause attachment preference of Persian EFL learners: Does segmentation play any role? *Journal of Psycholinguistic Research*, *51*, 683–706. https://doi.org/10.1007/s10936-021-09825-9

Malone, A., & Mauner, G. (2020). *Syntactic adaptation for reduced relative clauses is not reducible to task adaptation* [Poster presentation]. Presented at the 33rd annual CUNY Conference, Davis, California, USA. https://osf.io/wq2g7/download

Mardani, M., & Modarres, M. (2023). The effects of age, presentation mode (online, offline), and segmentation of ambiguous sentences on attachment preferences of female EFL learners. *European Journal of English Language Teaching*, *8*(1). http://doi.org/10.46827/ejel.v8i1.4695

Marefat, H., & Abdollahnejad, E. (2014). Acquisition of English relative clauses by adult Persian learners: Focus on resumptive pronouns. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *32*(4), 19-40. https://doi.org/10.22099/jtls.2014.1858

Marefat, H., & Farzizadeh, B. (2018). Relative clause ambiguity resolution in L1 and L2: Are processing strategies transferred? *Iranian Journal of Applied Linguistics (IJAL)*, *21*(1), 125-161.

Marefat, H., Samadi, E., & Yaseri, M. (2015). Semantic priming effect on relative clause attachment ambiguity resolution in L2. *Applied Research on English Language*, *4*(2), 78-95. https://doi.org/10.22108/are.2015.15504

Marinis, T. (2010). Using on-line processing methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 139-162). John Benjamins.

Maroof, D. A. (2012). *Statistical methods in neuropsychology: Common procedures made comprehensible*. Springer.

Marsden, E. (2020). Methodological transparency and its consequences for the quality and scope of research. In J. McKinley & H. Rose (Eds.), *The Routledge handbook of research methods in applied linguistics* (pp. 15-28). Routledge.

Marsden, E., Plonsky, L., Gudmestad, A., & Edmonds, A. (2018). Data, open science, and methodological reform in second language acquisition research. In C. G. Mayo (Ed.), *Critical reflections on data in second language acquisition* (pp. 219-228). Multilingual Matters.

Marsden, E., Thompson, S., & Plonsky, L. (2017). Open science in second language acquisition research: The IRIS repository of research materials and data. In C. Granget, M-A. Dat, D. Guedat-Bittighof- fer, & C. Cuet (Eds.), *Connaissances et Usages en L2/Knowledge and usage in L2: SHS web of conferences* (Vol. 38). https://doi.org/10.1051/shsconf/20173800013

Marsden, E., Thompson, S., & Plonsky, L. (2018). A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics*, *39*(5), 861-904. https://doi.org/10.1017/S0142716418000036

Matić, A., & Kovačević, M. (2022). Challenges of Different Approaches and Methodologies in Psycholinguistics: The Example of an RC Attachment Preference Study in Croatian. In J. Gervain, G. Csibra, K. Kovács (Eds.), *A Life in Cognition: Studies in Cognitive Science in Honor of Csaba Pléh* (pp. 125-136). Springer.

Mellor, D., Vazire, S., & Lindsay, D. S. (2018). Transparent science: A more credible, reproducible, and publishable way to do science. In R. J. Sternberg (Ed.), *Guide to publishing in psychology journals* (pp. 219-237). Cambridge University Press.

Melnyk, B. M., & Morrison-Beedy, D. (2012). *Intervention research: Designing, conducting, analyzing, and funding*. Springer.

Mendelsohn, A., & Pearlmutter, N. J. (1999). *Individual differences in relative clause attachment ambiguities* [Poster presentation]. Presented at the 12th Annual CUNY Conference on Human Sentence Processing, New York City, USA.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., Glennerster, R., Green, D. P., Humphreys, M., Imbens, G., Laitin, D., Madon, T. Nelson, L., Nosek, B. A., Petersen, M., Sedlmayr, R., Simmons, J. P., Simonsohn, U., Van der Laan. M. (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30-31. https://doi.org/10.1126/science.1245317

Miller, S. D., & Smith, D. E. P. (1985). Differences in literal and inferential comprehension after reading orally and silently. *Journal of Educational Psychology*, *77*(3), 341–348.

Miller, S. D., & Smith, D. E. P. (1990). Relations among oral reading, silent reading and listening comprehension of students at differing competency levels. *Reading Research and Instruction*, *29*(2), 73–84. https://doi.org/10.1080/19388079009558006

Miyao, M., & Omaki, A. (2006). *No ambiguity about it: Korean learners of Japanese have a clear attachment preference* [Paper presentation]. Presented at the 30th Annual Boston University Conference on Language Development. Boston, USA.

Moon, N., & Yun, H. (2021). The Role of Honorific Agreement in the Resolution of Relative Clause Attachment Ambiguity. *영어영문학*, *26*(4), 23-53. https://doi.org/10.46449/MJELL.2021.11.26.4.23

Morea, N., & Ghanbar, H. (2024). Q methodology in applied linguistics: A systematic research synthesis. *System*, *120*, 103194. https://doi.org/10.1016/j.system.2023.103194

Nakanishi, T. (2015). A meta-analysis of extensive reading research. *TESOL Quarterly*, *49*(1), 6-37. https://doi.org/10.1002/tesq.157

Nakano, Y. (2009). The influence of proficiency levels on resolving ambiguous relative-clause attachments in German as a second language. *言語と文化*, (12), 55-69.

Nestor, P. G., & Schutt, R. K. (2018). *Research methods in psychology: Investigating human behavior*. Sage Publications.

Norouzian, R. (2021). Interrater reliability in second language meta-analyses: The case of categorical moderators. *Studies in Second Language Acquisition*, *43*(4), 896-915. http://doi.org/10.1017/S0272263121000061

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719-748. https://doi.org/10.1146/annurev-psych-020821-114157

O'Brien, B. A., Wallot, S., Haussmann, A., & Kloos, H. (2014). Using complexity metrics to assess silent reading fluency: A cross-sectional study comparing oral and silent reading. *Scientific Studies of Reading*, *18*(4), 235-254. http://doi.org/10.1080/10888438.2013.862248

Omaki, A. (2005). *Working memory and relative clause attachment in first and second language processing* [Unpublished master's thesis]. University of Hawaii.

Øby, E. (2024). Assessing transparency and methodological precision in variable measurement within organizational research: implications for validity. *Quality & Quantity*, 1-18. https://doi.org/10.1007/s11135-024-01991-x

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning* (2nd ed.). Sage Publications.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295

Papadopoulou, D. (2006). *Cross-linguistic variation in sentence processing: Evidence from RC attachment preferences in Greek*. Springer.

Papadopoulou, D., & Clahsen, H. (2003a). Parsing strategies in L1 and L2 sentence processing: A study of relative clause attachment in Greek. *Studies in Second Language Acquisition*, *25*(4), 501-528. https://doi.org/10.1017/S0272263103000214

Papadopoulou, D., & Clahsen, H. (2003b). The role of lexical and contextual information in parsing ambiguous sentences in Greek. Working Paper. Essex Research Reports in Linguistics, University of Essex, Colchester, UK. https://www.researchgate.net/publication/228362839

Pigott, T. (2012). *Advances in meta-analysis*. Springer Science & Business Media.

Plonsky, L. (2013). Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition*, *35*(4), 655-687. https://doi.org/10.1017/S0272263113000399

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *The Modern Language Journal*, *98*(1), 450-470. https://doi.org/10.1111/j.1540-4781.2014.12058.x

Plonsky, L., & Gonulal, T. (2015). Methodological synthesis in quantitative L2 research: A review of reviews and a case study of exploratory factor analysis. *Language Learning*, *65*(S1), 9-36. https://doi.org/10.1111/lang.12111

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics*, *36*, 73-97. https://doi.org/10.1017/S0267190516000015

Plonsky, L., & Oswald, F. L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106-128). Routledge.

Plonsky, L., Hu, Y., Sudina, E., & Oswald, F. L. (2023). Advancing Meta-Analytic Methods in L2 Research. In A. Mackey, S. M. Gass (Eds.), *Current Approaches in Second Language Acquisition Research: A Practical Guide* (pp. 304-333). John Wiley & Sons.

Plonsky, L., Marsden, E., Crowther, D., Gass, S. M., & Spinner, P. (2020). A methodological synthesis and meta-analysis of judgment tasks in second language research. *Second Language Research, 36*(4), 583-621. https://doi.org/10.1177/0267658319828413

Prasad, G., & Linzen, T. (2021). Rapid syntactic adaptation in self-paced reading: Detectable, but only with many participants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://doi.org/10.1037/xlm0001046

Price, K. W., Meisinger, E. B., Louwerse, M. M., & D'Mello, S. (2016). The contributions of oral and silent reading fluency to reading comprehension. *Reading Psychology, 37*(2), 1-35. https://doi.org/10.1080/02702711.2015.1025118

Prior, S. M., Fenwick, K. D., Saunders, K. S., Ouellette, R., O'Quinn, C., & Harvey, S. (2011). Comprehension after oral and silent reading: Does grade level matter?. *Literacy Research and Instruction*, *50*(3), 183-194. https://doi.org/10.1080/19388071.2010.497202

Pynte, J., & Colonna, S. (2001). Competition between primary and non-primary relations during sentence comprehension. *Journal of Psycholinguistic Research, 30*, 569-599. https://doi.org/10.1023/A:1014278905819

Rah, A. (2009). *Sentence processing in a second language: Ambiguity resolution in German learners of English* [Unpublished doctoral dissertation]. Universität zu Köln.

Reifegerste, J., Jarvis, R., & Felser, C. (2020). Effects of chronological age on native and nonnative sentence processing: Evidence from subject-verb agreement in German. *Journal of Memory and Language*, *111*, 104083. https://doi.org/10.1016/j.jml.2019.104083

Riazi, A. M., & Amini Farsani, M. (2024). Mixed-methods research in applied linguistics: Charting the progress through the second decade of the twenty-first century. *Language Teaching, 57*(2), 143-182. https://doi.org/10.1017/S0261444823000332

Robinson, M. F., Meisinger, E. B., & Joyner, R. E. (2019). The influence of oral versus silent reading on reading comprehension in students with reading disabilities. *Learning Disability Quarterly*, *42*(2), 105-116. https://doi.org/10.1177/0731948718806665

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638. https://doi.org/10.1037/0033-2909.86.3.638

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research, and Evaluation*, *11*(1), 10. https://doi.org/10.7275/9wph-vv65

Samadi, E., Karimi, M. N., & Babaii, E. (2022). The role of semantic priming in relative clause attachment ambiguity resolution in Persian. *Journal of Language Horizons*, *6*(1), 247-261. https://journal.alzahra.ac.ir/article_5221.html

Scheepers, C., Raffray, C. N., & Myachykov, A. (2017). The lexical boost effect is not diagnostic of lexically-specific syntactic representations. *Journal of Memory and Language*, *95*, 102-115. https://doi.org/10.1016/j.jml.2017.03.001

Schimmel, N., & Ness, M. (2017). The effects of oral and silent reading on reading comprehension. *Reading Psychology*, *38*(4), 390–416. https://doi.org/10.1080/02702711.2016.1278

Siriwittayakorn, T., Miyamoto, E. T., Ratitamkul, T., & Cho, H. (2014, December). *A Non-local Attachment Preference in the Production and Comprehension of Thai Relative Clauses* [Paper presentation]. Presented at the 28th Pacific Asia Conference on Language, Information and Computing (pp. 575-584). https://aclanthology.org/Y14-1066.pdf

Siriwittayakorn, T., Miyamoto, E. T., & Ratitamkul, T. (2015). *Contextual effects and locality preferences in relative clause attachment in Thai* [Paper presentation]. Presented at the EuroAsianPacific joint conference on cognitive science, Torino, Italy.

Slaghuis, W. L., Lovegrove, W. J., & Davidson, J. A. (1993). Visual and language processing deficits are concurrent in dyslexia. *Cortex*, *29*(4), 601-615. https://doi.org/10.1016/S0010-9452(13)80284-5

Soares, S. M. P., Kupisch, T., & Rothman, J. (2022). Testing potential transfer effects in heritage and adult L2 bilinguals acquiring a mini grammar as an additional language: An ERP approach. *Brain Sciences*, *12*(5), 669. https://doi.org/10.3390/brainsci12050669

Sokolova, M., & Slabakova, R. (2019). L3 Sentence Processing: Language-Specific or Phenomenon-Sensitive?. *Languages*, *4*(3), 54. https://doi.org/10.3390/languages4030054

Sokolova, M., & Slabakova, R. (2021). Processing similarities between native speakers and non-balanced bilinguals. *International Journal of Bilingualism*, *25*(6), 1655-1679. https://doi.org/10.1177/13670069211033647

Son, M. (2020). *Subject-verb agreement in written English by L1 Norwegian university students: Error patterns, causes, and implication for teaching* [Unpublished master's thesis]. UiT The Arctic University of Norway. https://munin.uit.no/bitstream/handle/10037/20511/thesis.pdf

Spitschan, M., Schmidt, M. H., & Blume, C. (2020). Principles of open, transparent and reproducible science in author guidelines of sleep research and chronobiology journals. *Wellcome Open Research*, *5*. https://doi.org/10.12688/wellcomeopenres.16111.2

Stetie, N. A., & Zunino, G. M. (2021). Revisiting attachment preferences in Spanish: is there a high attachment bias? [Poster presentation]. Presented at the 34th CUNY Conference on Human Sentence Processing. University of Pennsylvania. USA.

Swets, B., Desmet, T., Hambrick, D. Z., & Ferreira, F. (2007). The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General*, *136*(1), 64. https://doi.org/10.1037/0096-3445.136.1.64

Tan, M., & Foltz, A. (2020). Task sensitivity in L2 English speakers' syntactic processing: Evidence for Good-Enough processing in self-paced reading. *Frontiers in Psychology*, *11*, 575847. https://doi.org/10.3389/fpsyg.2020.575847

Traxler, M. J., & Tooley, K. M. (2007). Lexical mediation and context effects in sentence processing. *Brain Research*, *1146*, 59-74. https://doi.org/10.1016/j.brainres.2006.10.010

Urdan, T. C. (2022). *Statistics in plain English*. Routledge.

Vafaee Seresht, K., & Marefat, H. (2022). Methodological synthesis of working memory capacity measures in relative clause attachment ambiguity resolution studies. *Teaching English as a Second Language Quarterly (Formerly Journal of Teaching Language Skills)*, *41*(4), 113-172. https://doi.org/10.22099/tesl.2022.42703.3084

van Gompel, R. P. G., Wakeford, L. J., & Kantola, L. (2022). No looking back: The effects of visual cues on the lexical boost in structural priming. *Language, Cognition and Neuroscience*, *37*(1), 1-10. https://doi.org/10.1080/23273798.2022.2036782

Vannoy, S. A., Medlin, B. D., & Chen, C. C. (2011). Enhancing the trust of members in online social networks: An integrative technical and marketing perspective. *International Journal of Virtual Communities and Social Networking (IJVCSN)*, *3*(4), 15-31. http://doi.org/10.4018/978-1-4666-4022-1.ch014

Wang, S. V., Sreedhara, S. K., & Schneeweiss, S. (2022). Reproducibility of real-world evidence studies using clinical practice data to inform regulatory and coverage decisions. *Nature communications*, *13*(1), 5126. https://doi.org/10.1038/s41467-022-32310-3

Weiner, M. J. (1980). The effect of incentive and control over outcomes upon intrinsic motivation and performance. *The Journal of Social Psychology*, *112*(2), 247-254.

Yao, B., & Scheepers, C. (2018). Direct speech quotations promote low relative-clause attachment in silent reading of English. *Cognition*, *176*, 248-254. https://doi.org/10.1016/j.cognition.2018.03.017

Zedeck, S. (2014). *APA dictionary of statistics and research methods*. American Psychological Association.

Zhang, M., & Plonsky, L. (2020). Collaborative writing in face-to-face settings: A substantive and methodological review. *Journal of Second Language Writing*, *49*, 100753.

Ziegler, N. (2016). Synchronous computer-mediated communication and interaction: A meta-analysis. *Studies in Second Language Acquisition*, *38*(3), 553-586. https://doi.org/10.1017/S027226311500025X

Zogmaister, C., Vezzoli, M., Facchin, A., Conte, F.P., Rizzi, E., Giaquinto, F., Cavicchiolo, E., Fusco, G., Pegoraro, S. and Simioni, M. (2024). Assessing the Transparency of Methods in Scientific Reporting. *Collabra: Psychology*, *10*(1). https://doi.org/10.1525/collabra.121243