A Methodological Synthesis of Judgment Tasks in Second Language Research

Journal:	*Second Language Research*
Manuscript ID	SLR-18-0029.R1
Manuscript Type:	Original Article
Keywords:	methodological synthesis, judgment task, instrumentation, meta-analysis, L2 knowledge
Abstract:	Judgment tasks (JTs, often called acceptability or grammaticality judgment tasks) are found extensively throughout the history of second language (L2) research (Chaudron, 1983). Data from such instruments have been used to investigate a range of hypotheses and phenomena, from generativist theories to instructional effectiveness. Though popular and convenient, JTs have engendered considerable controversy, with concerns often centered on their construct validity in terms of the type of representations they elicit, such as implicit or explicit knowledge (Ellis, 2005; Vafaee et al., 2016). A number of studies have also examined the impact of JT conditions such as timed vs. untimed, oral vs. written (e.g., Murphy, 1997; Spada et al., 2015). This paper presents a synthesis of the use of JTs and a meta-analysis of the effects of task conditions on learner performance. Following a comprehensive search, 385 JTs were found in 302 individual studies. Each report was coded for features related to study design as well as methodological, procedural, and psychometric properties of the JTs. These data were synthesized in order to understand how this type of instrument has been implemented and reported. In addition to observing a steady increase in the use of JTs over the last four decades, we also found many of the features of JTs, when reported, varied substantially across studies. In terms of the impact of JT design, whereas modality was not found to have a strong or stable effect on learner performance (median d=.14; IQR=1.04), scores on untimed JTs tended to be substantially higher than when timed (d=1.35; IQR=1.74). In examining these features and their links to findings, this paper builds on a growing body of methodological syntheses of L2 research instrumentation (e.g., Derrick, 2016; Marsden et al., in press) and makes a number of empirically grounded recommendations for future studies involving JTs.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

SCHOLARONE™
Manuscripts

A methodological synthesis of judgment tasks in second language research

To understand human language, we must understand the grammatical boundaries of language. Since the early days of generative linguistic research, one of the most common ways of determining those boundaries has been to use judgments of grammaticality, or to use a more appropriate term, judgments of acceptability. However, in many disciplines, including second language (L2) research both within and outside of the generative tradition, the validity, reliability, and suitability of acceptability judgments for data elicitation has been questioned. In particular, there are longstanding concerns about the type of knowledge that judgment data reflect. Additionally, particularly in L2 research, there are numerous variables and issues that need to be considered not only in relation to data elicitation and test format, but also in relation to reporting, analysis, and interpretation. In this article, we present a methodological synthesis of 385 judgment tasks (henceforth JTs) found in 302 individual studies. In doing so, we systematically examine the use of JTs and the ways JTs have been designed, administered and reported. We also quantify the influence of JT design on observed L2 knowledge in this body of research.

Background

As Sprouse (2013) noted, "[a]cceptability judgments form a substantial portion of the empirical foundation of nearly every area of linguistics…and nearly every type of linguistic theory" (p. 1). In L2 research, JTs have been used (a) across a wide variety of theoretical frameworks and substantive domains, including generative and cognitive psycholinguistic studies, L2 pragmatics, and classroom-based research, and (b) for numerous and wide-ranging purposes, including to show development over time, to assess proficiency, to screen participants, and to determine knowledge types (see Spinner and Gass, in press, for greater detail and

elaboration on the history of acceptability judgment tasks in linguistics and L2 research). To some extent, the widespread use of JTs can be attributed to the influence of linguistic research methods on L2 research, but another important factor is their practicality. For example, it has been widely thought that JTs are relatively easy to develop and administer; their scoring is relatively straightforward; they can target language features that are rare or difficult to elicit in production (Loewen, 2009); and because of apparent commonalities in target structures, formats, and the overarching ambition to elicit judgments about language, they allow for cross-study comparisons.

Despite the widespread use of JTs, there have been and continue to be concerns about important methodological issues that affect our ability to interpret what judgment data mean (see Chaudron, 1983, for an early review of the use of JTs in L1 and L2 acquisition). JTs have even been challenged as a valid research tool for understanding acquisition. Perhaps the earliest and most direct challenge for L2 researchers came from Selinker (1972), who said that researchers should "focus our analytical attention upon *the only observable data to which we can relate theoretical predictions:* the utterances which are produced when the learner attempts to say sentences of a TL" (target language) (pp. 213-214, emphasis in original). This notion was furthered by Selinker (1974), who noted that "the only observable data from meaningful performance situations we can establish as relevant…are…IL utterances produced by the learner" (p. 35). In other words, in his view, only oral production data are valid data—which rules out the usefulness of judgment data for L2 research. However, Schachter, Tyson, and Diffley (1976), among others, challenged Selinker's belief that judgment data are inappropriate or unnecessary. In particular, a focus on production data alone makes it impossible to determine what learners do not know or cannot do and to measure their sensitivity to norms in the input.

(See Gass and Polio, 2014, for a more detailed discussion of this issue). Despite these concerns and debates about judgment data in the early years of the field, judgment data have continued to be questioned throughout the years, with justification often being given for using them. We turn to a discussion of more recent debates about the nature of judgment data.

The nature of knowledge elicited by JTs

In their most common form, judgments are elicited on isolated sentences using either a dichotomous response or a Likert-style response. One of the main concerns about JTs is such a task is 'unnatural'. That is, they do not reflect real-world use of language. Similarly, there is concern that acceptability judgments may not tap into learners' implicit knowledge of language, but rather only access explicit or even only metalinguistic knowledge. For some researchers, measuring explicit or metalinguistic knowledge is exactly the goal, but often researchers are attempting to discern the nature of implicit knowledge underlying an interlanguage grammar. Implicit knowledge is often considered a type of knowledge that is most useful to understanding the nature of L2 knowledge in that it is more stable, less prone to decay (e.g., after instruction), more immediately accessible, and less context-dependent (Rebuschat and Williams, 2012). There is some debate about whether acceptability judgments can tap into more implicit knowledge types, and if so, how they should be designed and analyzed to best accomplish this task (see e.g., Ellis, 2005).

As part of this debate, various task features and conditions have been tested as a means to validate measures of different types of L2 knowledge elicited by JTs. For instance, using timed or speeded judgments may lead learners to employ (more) implicit knowledge in part because there is not enough time to think back to learned rules, typical of those presented in a classroom

context (Ellis, 2005). Godfroid et al. (2015) used eye-tracking to monitor learners' eye

movements during timed and untimed judgement tasks and found that learners regressed more

during the untimed tasks; they concluded that the untimed tasks may lead to more use of explicit

knowledge than the timed tasks. However, the matter is far from settled. Kim and Nam (2016)

compared timed judgment data and oral elicited imitation data and found that the latter to be a

better indicator of implicit knowledge. The authors also found that JTs measured implicit

knowledge more directly when the stimuli were presented aurally as opposed to in the written

modality.

Gutiérrez (2013), investigating the construct validity of JTs, administered a timed and an

untimed JT and a metalinguistic knowledge test to 49 L2 learners of Spanish. The timed JT was

intended to measure implicit knowledge and the other two were intended to measure explicit

knowledge. Using exploratory and confirmatory factor analyses, grammatical sentences (timed

and untimed JTs) loaded on a construct interpreted as implicit knowledge and the ungrammatical

sentences loaded on a constructed interpreted as explicit knowledge. Based on their results, the

author argued that time pressure does not distinguish between knowledge types; what does

distinguish between knowledge types is whether or not a sentence is grammatical or

ungrammatical.

Vafaee, Suzuki, and Kachisnke (2017) argued against Gutiérrez's (2013) heavy reliance

on JTs as the means to measure explicit and implicit knowledge. They conducted a confirmatory

factor analysis based on data elicited from 79 Chinese L1 speakers learning English as an L2.

They used timed and untimed grammaticality JTs, a self-paced reading task, a word-monitoring

task, and a metalinguistic task to test the extent to which a JT reflected implicit knowledge. They

concluded that grammaticality judgments, regardless of being timed or untimed, were not a

suitable elicitation technique for measuring implicit knowledge.

Other design and analysis issues may also influence learners' use of implicit or explicit knowledge or may, more generally, influence participant scores on JTs. For instance, having learners indicate errors or provide corrections to them while they make judgments may lead to increased use of explicit and/or metalinguistic knowledge (e.g., Ellis, 1991). Additionally, task modality (e.g., written versus aural), response types (e.g., binary vs. scaled), the balance of grammatical/acceptable to ungrammatical/unacceptable items, and confidence ratings, all of which feature to varying degrees in JTs, may also impact learner performance.

Framing their study within the context of the frequent use of JTs, Shiu, Yalçın, and Spada (2018) pointed out that "it is essential to continue to explore how design features contribute to leaners' GJT performance (p. 216). They considered four key design features (time constraints, task stimulus, task modality, and target features) and investigated the potential outcomes of results when these features were manipulated. They found that time constraints, modality, and task stimulus impacted results, thereby further illustrating the complexities involved in using and interpreting judgment data.

In another study focusing on design features, Murphy (1997) investigated declarative sentences with embedded questions and *wh*-questions that violated Subjacency, asking whether there might be differences in outcomes if sentences were presented visually or aurally. Participants were not only slower when sentences were presented aurally, but even more important they were less accurate. She emphasized the importance of methodology when interpreting research results.

The studies reviewed here underscore the overall importance of a thorough understanding of the potential impact of different design features.

Methodological transparency and rigor

In recent years, methodological issues have been a focus of attention as the field of L2 research matures. Emphasizing this point, Byrnes (2013: 825) points out that "methodological issues inherently merit a certain level of attention inasmuch as they assure the quality of our work. But it appears that at this point in the development of applied linguistics, they demand a kind of professional scrutiny that goes directly to the core of what we do and what we know and what we can tell our publics that we know—and not only how we do it."

As the result of greater attention as well as, perhaps, improved training and stricter journal standards (e.g., Norris, Plonsky, Ross, & Schoonen, 2015), methodological rigor in the field appears to be improving. However, there is still a long way to go: Using techniques characteristic of methodological synthesis, Derrick (2016) reviewed the reporting practices surrounding 925 data collection instruments found in 385 articles spanning a wide variety of domains in L2 research. Her results indicate a general lack of transparency and rigor in data collection, reporting, and methodological transparency (as discussed by Marsden and Plonsky, 2018). For example, only 17% of the sample was found to have reported piloting their data collection instruments, and reliability coefficients were available for only 28% of the sample (see similar results found by Plonsky, 2013).

Of course, we cannot equate unreported reliability with low reliability (i.e., high measurement error). Plonsky and Derrick (2016) set out to measure the error in L2 research instruments as reported in primary studies. Their approach, known as 'reliability generalization meta-analysis' (Rodriguez and Meada, 2006; Wheeler, Vassar, Worley, and Barnes, 2011), involves systematically reviewing and aggregating reliability coefficients to estimate the overall reliability in a given sample of studies. The median reliability coefficient (for internal

consistency) was .82, with somewhat higher interrater and intrarater estimates. This estimate was also found to vary as a function of several additional features of L2 designs and instruments.

Marsden, Thompson, and Plonsky (2018b) represents another recent methodological synthesis with parallels to the present study. Their review, like ours, focused on the use of one particular elicitation procedure, self-paced reading, which has likewise been the subject of a number of debates concerning construct validity as well as what might constitute appropriate implementation and design (e.g., Keating and Jegerski, 2015; Roberts, 2016). Marsden et al.'s results indicated that the features of self-paced reading tasks vary widely. For example, there was little consistency in terms of target regions being analyzed, even among reports examining the same linguistic features. Substantial variability was also observed in a number of other features such as the inclusion of comprehension questions, the number of target vs. distractor or filler items, and sentence length, all potentially affecting learner outcomes as measured in milliseconds. Reporting of data related to self-paced reading was also inconsistent. For example, only two of the 64 studies in their sample provided an estimate of instrument reliability.

Beyond the reviews described here, the current study follows on the heels of a growing body of methodological investigation and meta-science. Some syntheses in this area deal with issues particular to a given substantive domain (e.g., L2 written feedback, Liu and Brown, 2015; interactionist SLA, Plonsky and Gass, 2011; learner corpus research, Paquot and Plonsky, 2017; and interaction and computer-mediated communication, Ziegler, 2016). Others, however, address one or more practices across domains, such as: mixed methods (Hashemi and Babaii, 2013); reporting practices (Larson-Hall and Plonsky, 2015); replication research (Marsden, Morgan-Short, Thompson, and Abugaber, 2018a); quantitative analyses and reporting practices, (Plonsky, 2013, 2014); study designs (Marsden and Torgerson, 2012; Plonsky and Gonulal, 2015);

multiple regression (Plonsky & Ghanbar, in press; Plonsky and Oswald, 2017); and structural

equation modeling (Winke, 2014). Motivating this line of research is, in part, an acute awareness

that "respect for the field […] can come only through sound scientific progress" (Gass, Fleck,

Leder, and Svetics, 1998: 407). However, an even more fundamental rationale for describing and

evaluating the empirical efforts in L2 research is the need to ensure the validity of the theoretical

and practical claims we make based on our research. In view of the importance of this for the

progress and usefulness of science both generally and in the case of efforts in the field of L2

research, we support the notion put forth by Plonsky (2013) that methodological practices and

study quality need to be measured, not assumed. Following this line of thinking, in the present

study, we systematically review the methodological issues and choices involved when eliciting

judgment data in L2 research and we investigate how some of these choices can affect

substantive findings.

Research questions

1. To what extent and in what ways have JTs been used in L2 research?

2. How have JTs been designed and administered in L2 research?

3. To what extent do design features of JT (modality and timing) influence the scores they elicit?

Method

Study Identification

        The first step in conducting a research synthesis or meta-analysis is to identify

inclusion/exclusion criteria based on the substantive or methodological domain of interest

(Plonsky and Oswald, 2015). As our primary criterion for inclusion in the present review, a study

had to include one (or more) JTs that targeted participants' morphosyntactic knowledge in their

L2 (L3, L4, etc.). Studies targeting either L1 morphosyntactic knowledge or L2 pragmatic or

discourse knowledge were excluded in order to arrive at a more homogenous sample. In addition, specific criteria were established regarding the type of JTs eligible for inclusion, chosen in order to arrive at a relatively homogeneous and representative—if not exhaustive—sample of the use of JTs in L2 research:

- Outcome measures had to be accuracy (i.e., based on a judgment) rather than a measure of processing (e.g., ERP, reaction times).

- Judgment was made on an individual sentence (rather than on a longer stretch of discourse).

Studies with items requiring a comparison of grammaticality/acceptability across multiple sentences (e.g., which sentence is most acceptable?) were excluded. Also excluded were magnitude estimation tasks.

With the eligibility criteria established, we began an extensive keyword database search. Following Plonsky and Oswald (2015) we included Linguistics and Language Behavior Abstracts (LLBA), Education Resources Information Center (ERIC), PsycInfo, and Google Scholar. In addition, we searched the L2 Research Corpus (Plonsky, n.d.) as well as the IRIS Digital Repository (https://www.iris-database.org; see Marsden, Mackey, and Plonsky, 2016). Although it was likely that there would be significant overlap between the six databases, comprehensiveness was prioritized over redundancy (following Lee et al., 2015; Plonsky and Oswald, 2015). The key words were 'grammaticality', 'acceptability', and 'truth value', which were required to occur alongside both 'judgment' (or 'judgement') and either 'task' or 'test' (e.g., 'acceptability judgment test', 'grammaticality judgment test'). A number of additional criteria were then also applied to reduce the number of identified studies into a more homogeneous sample. First, all included studies were required to feature primary research

published in a journal on the Social Science Citation Index (SSCI) (http://ip-science.thomsonreuters.com/mjl/), with all non-indexed journal articles, book chapters, dissertations, theses, and conference proceedings excluded. Any duplicate reports or datasets (i.e., data that had been analyzed exactly as they were analyzed in previous publications) were excluded. In total, 302 studies were identified comprising a total of 385 JTs (see materials associated with this study on IRIS; iris-database.org). (Just under a third of the full JTs and records for all the articles in our sample are available under 'Special Collections' at www.iris-database.org, discussed below).

Coding

A coding scheme was designed to collect data on each study and on the JTs in our sample. This tool was similar in design to those of other methodological syntheses (e.g., Plonsky, 2013; Marsden et al., 2018b). The final version included 69 items eliciting substantive and methodological features across five themes: (a) *Bibliographic Information* (e.g., authors, journal, year); (b) *Study Design* (e.g., *N*, target L2, one-shot vs. developmental); (c) *JT Design* (timed vs. untimed; modality; number of items; balance between grammatical/acceptable vs. ungrammatical/unacceptable items); (d) *Reliability, Transparency, and Validity* (e.g., reporting of instrument reliability; instrument availability; and (e) *Theoretical Framework* the study was conducted under (e.g., formal, functional, usage-based). Appendix A lists all coded features along with descriptors used. In addition, the full instrument in its original format (an Excel spreadsheet), including the entire dataset for the present study will be available on the IRIS database upon publication.

Following several rounds of revising and piloting the instrument, one of the study's co-authors coded the entire sample. This co-author then provided extensive training to a research

assistant (a doctoral student in applied linguistics) who re-coded 15% of the JTs in the sample (K_{study} = 43, K_{task} = 62). Though a low percentage of the overall sample, the total number of recoded studies is well above the minimum requirement by Lipsey and Wilson (2001). Although agreement between the first and second coder was generally high (median across all items = 87%), with an overall Cohen's Kappa (κ) of .72, the range of reliability coefficients (Kappa) across categories was wide (κ = .11–1.00). The two coders discussed and resolved disagreements whenever present. Based on discussions amongst the researchers, problematic studies were recoded. The source of low Kappa-values was almost always attributed to a difference between items coded as "No" (i.e., "not present") and "Not Reported". For example, when asked if the judgment task in question had two counterbalanced forms, one coder often recorded "No" (no counterbalancing), while the other would say that this information was "Not Reported". While the range of Kappa values is not ideal, the overall Kappa of .72 is within range of many other meta-analyses and methodological syntheses (e.g., Marsden et al., 2018b; Paquot and Plonsky, 2017). In addition to the coded features, Appendix A reports both Kappa and percent agreement for each item in the coding scheme.

Analysis

RQ1 was concerned with the frequency, publication outlet, context, and design of studies employing JTs in L2 research. To address this question, we calculated the frequencies and percentages of JTs found overall, over time, and across different journals. We also calculated the frequency and percentage of major types of designs as well as demographic and contextual features found in our sample.

RQ2 focused on the design and implementation of the JTs themselves. Similar to RQ1, frequencies and percentages of JT features were calculated across the study sample. One of the

features of interest was reliability (internal measurement consistency). We recorded not only

whether estimates of reliability were available but also the estimates as reported. Using reliability

generalization meta-analysis, reliability estimates associated with the JTs in our sample were

aggregated, and measures of central tendency and dispersion were examined.

Our third and final research question addressed the extent to which different features of

JTs might impact the data they produce. To answer this question, we first identified the studies

that administered the same JT but with different design features to the same sample of

participants. Specifically, we selected those studies that investigated either or both of the

following two parameters within the same study: modality (aural vs. written) and/or timing

(timed vs. untimed). These variables were chosen for this phase of the analysis on the grounds

that they are (a) commonly discussed in the literature on JTs, (b) varied substantially across the

sample, and (c) were available in sufficient numbers for within-group comparisons. When

sufficient data were available (i.e., means, standard deviation, and sample size), we then

calculated a Cohen's *d* value for these contrasts (e.g., written vs. aural), representing the

standardized mean difference between the JT scores in each pair of conditions within each

parameter. The *d* values were then combined and compared using standard meta-analytic

procedures (Plonsky and Oswald, 2015).

Results

RQ1. To what extent and in what ways have JTs been used in L2 research?

Our first research question was concerned with when, how, and where JTs are used in L2

research. As a first step in understanding this particular tool, we examined the frequency of its

use over time and across different journals. As shown in Figure 1 and, in greater granularity in

the following Sparkline (see Larson-Hall, 2017;), the field's use of JTs has increased

substantially and regularly across the last 4 decades, with between approximately 40-50 more JT

studies published each decade since 1980. This regular increase is in part a reflection of the total

increase in volume of publications in the field. JTs are also found across a number of different

journals. However, Figure 2 shows that there is much stronger presence of JTs in some journals

(e.g., *Second Language Research*, *Studies in Second Language Acquisition*) than others (e.g.,

TESOL Quarterly), though of course this is partially influenced by the longevity of each
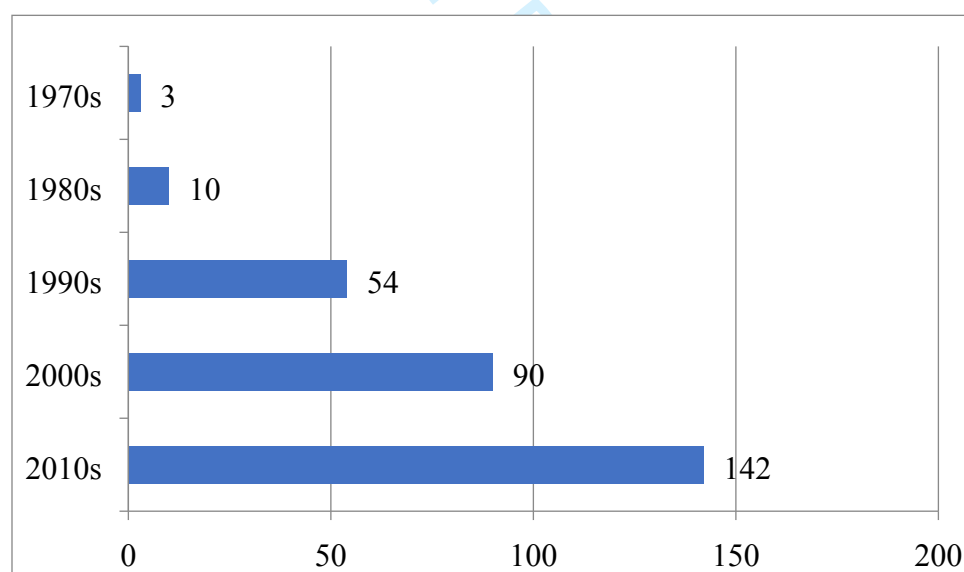
particular journal.

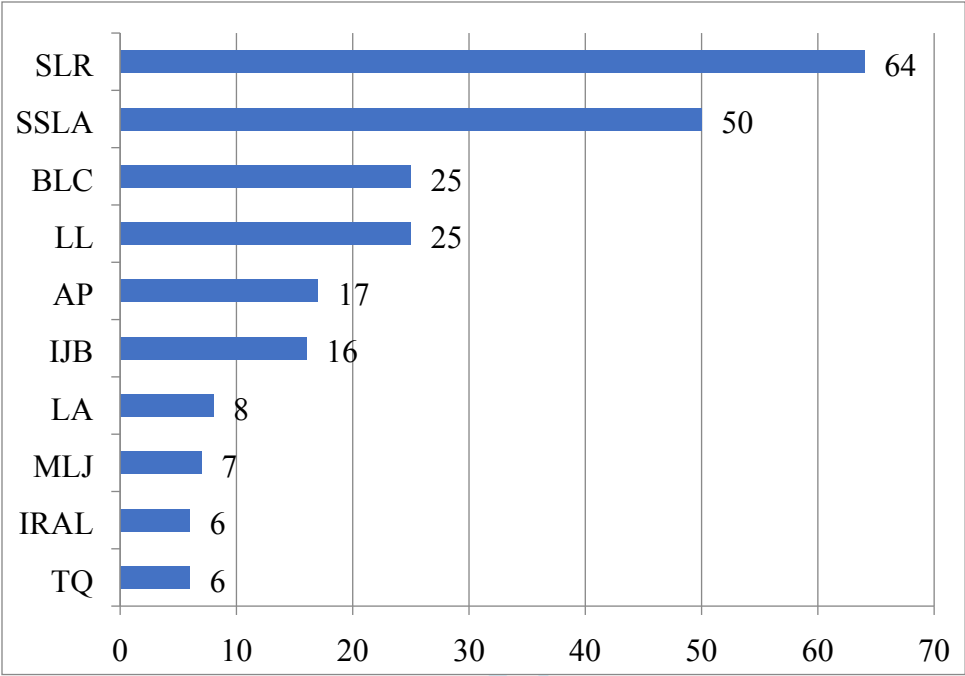Figure 1. Frequency of JTs Observed Over Time

Figure 2. Raw frequency (>5) of JTs Found in L2 Research Journals

Note. SLR = *Second Language Research*; SSLA = *Studies in Second Language Acquisition*; BLC = *Bilingualism: Language and Cognition*; LL = *Language Learning*; AP = *Applied Psycholinguistics*; IJB = *International Journal of Bilingualism*; LA = *Language Awareness*; MLJ = *Modern Language Journal*; IRAL = *International Review of Applied Linguistics*; TQ = *TESOL Quarterly*

Moving toward the types of research where JTs are found, approximately two-thirds (77%) of the studies in our sample employed a one-shot design. However, this does not mean that JTs have not been used in research investigating L2 development. It was not uncommon to find cross-sectional or 'pseudo-longitudinal' designs, with 24% of the sample comparing performance across different proficiency levels. Additionally, 60% of the sample compared learner responses to a native speaker group. Also related to design, we note that the native speaker samples (median $n = 20$) in this body of research were often much smaller than those of L2 learners (median $n = 47$), a choice likely associated with greater consistency generally

observed among native speakers and/or the greater difficulty of accessing native speaker

participants for some languages.

Study designs also varied in terms of the languages and learners involved. As we might

expect, a large portion of the sample involved English as an L1 (45%) or target language (59%).

Other L1s found with some regularity (i.e., ≥5%) included Chinese (15%), Japanese (9%),

French (8%), Arabic (6%), Spanish (6%), and Korean (5%). Target languages other than English

(≥5%) included Spanish (17%), French (11%), and Chinese (5%). There was a roughly equal

split between studies examining L2 knowledge among foreign (51%) as opposed to second

(41%) language learners. The remainder of the sample involved learners of artificial languages

(e.g., Ehrich and Meuter, 2009) or of real languages being learned by participants only for the

purpose of the study (e.g., Andringa and Curcic, 2015). Heritage language learners/users were

present in 7% of the sample. Primary studies were also coded for the proficiency of the learners

involved. Recognizing that authors did not often provide justification (such as qualifications or

amount of instruction) for claims regarding proficiency (17%), we found that 24% of the samples

were beginners, 33% intermediate learners, and 47% advanced learners. (Note: Total percentage

here adds to greater than 100% because some studies included groups at multiple proficiency

levels. We also note that proficiency levels were coded based on authors' labels of low,

intermediate, and advanced proficiency, when provided, and we recognize the potential for

inconsistencies across authors). 80% of the learners involved in this body of research were

adults; 4% were children and 3% were teens. The ages of participants in the remainder of the

studies were either mixed or not reported.

RQ2. How have JTs been designed and administered in L2 research?

One of the central interests of the current study was to understand the different features of

JTs found in published L2 research. Toward this end, we coded a number of features, the results

of which we summarize here and which, overall, can be seen to vary a great deal across the

sample.

One of the fundamental characteristics of any language test is the number of items it

includes. Although the median number of total items per instrument was 50 (median target items

= 32, IQR = 40), the minimum and maximum values range from 4 to 300, presenting in many

cases potential for participant fatigue. To offset this potential threat to internal validity, a small

number of studies (6%) reported offering participants a break. We note here that it is difficult to

determine precise item totals because 26% of the studies we surveyed were unclear on whether

distractor items were included and, if so, how many. One way to understand such variability in

the length of JTs is to consider the number of target features being tested. Although a clear

majority (66%) were focused on a single structure, individual JTs in our sample assessed learner

knowledge on up to 17 different structures. Across the entire sample, 33 unique target features

were identified with a fairly large median of 21 items per feature (interquartile range = 24).

Figure 3 displays the frequency of all those occurring in five or more studies. The most frequent

structure, word order, is also the broadest category. The remaining features represent a range of
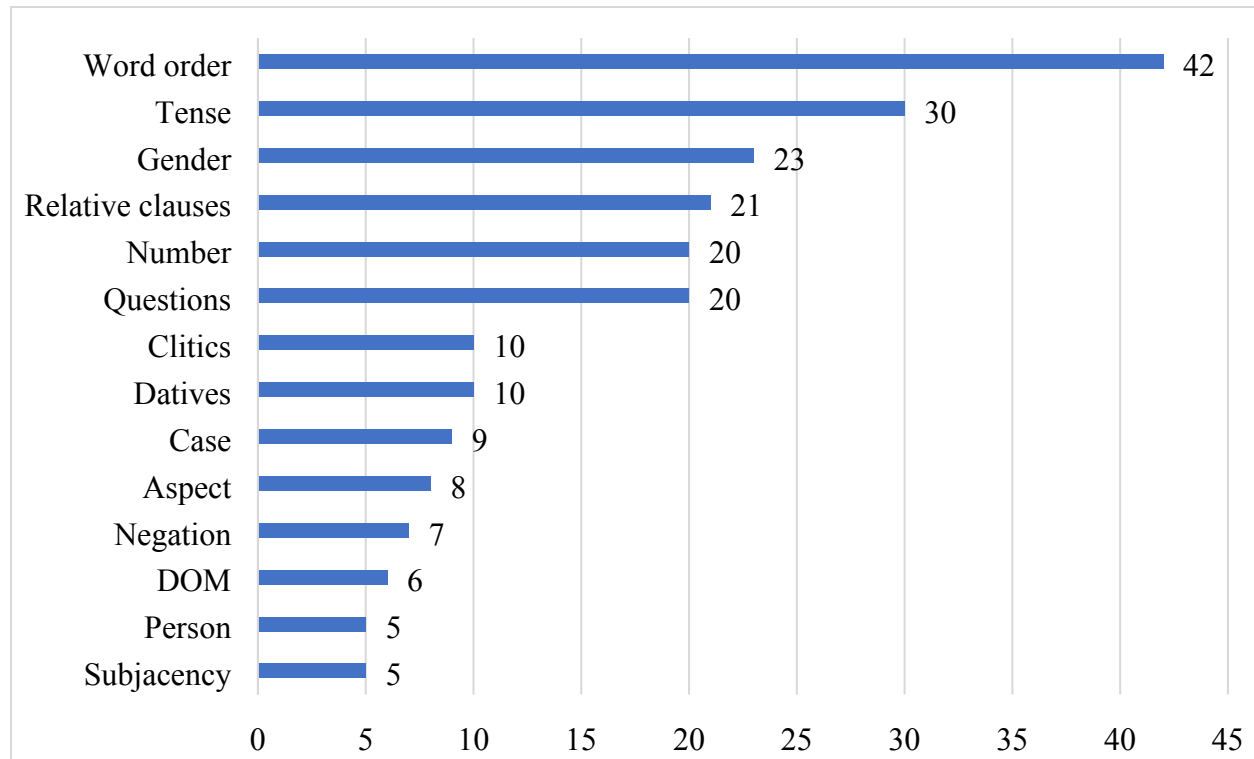
morphosyntactic targets.

Figure 3. Raw frequency of target features examined in JTs

Note. DOM = differential object marking

Beyond the number and focus of the features they assess, JTs vary along a number of additional features including test modality and format. Although it was not uncommon for JTs to be administered in the aural modality (16%), most JTs (58%) were administered as a written task, and 6% were provided in both modalities. Modality was not reported in an additional 18% of the sample. This feature is particularly critical given that one might expect learner performance to vary as a function of the modality in which the JT is administered, a point we return to in response to RQ3. A related matter is the format of the test. Approximately one-third of the JTs in our sample were administered in a paper-and-pencil (29%) or computer-based (36%) format; the format for the remaining studies was unreported.

In addition to testing format, another condition found to vary in JT administration is

17

timing. Whereas 55% of the sample did not report whether time restraints were involved, explicit

reference to a lack of a time limit was reported for 28% of the sample. The remaining 17% of the

studies imposed a time limit on participants to complete a judgement. However, justification for

the length of time given to participants to respond to an item was not uniformly provided: 33%

and 2% calculated the time limit based on native and non-native speaker norms (or piloting),

respectively. Loewen and Erlam (2006), for example, calculated the median time required by

native speakers to respond to each item in their study. They then added to this an additional 20%

"to allow for the slower processing speed of the L2 learners" (p. 8).

One broad category of JT features we explored is concerned with test and item

construction. It is often suggested that items be randomized to avoid drawing participants'

attention to particular structures being targeted (e.g., Chaudron, 1983). However, only 16% of

the sample reported to have done so. 21% did not randomize items, and 63% did not report

whether items were randomized or not. Compared to randomization, a slightly higher proportion

(26%) of studies was found to balance the number of ungrammatical/unacceptable versus

grammatical/acceptable items; 49% of the JTs did not do so, and 26% did not report whether or

not the JT was comprised of an equal number of grammatical/acceptable and

ungrammatical/unacceptable items. In some studies, contextual information was provided to

accompany JT items. This can, for example, help to coerce particular features by ensuring

specific meanings or functions are expressed, or it can increase the 'naturalness' of the test by

embedding the sentences in a meaningful context. In our study sample, 9% included a story, 8%

paired items with an image or video, and the items in 2% of the JTs were embedded within a

larger body of prose. In the remaining 81% of the JTs, no additional or contextual information

was provided.

We also examined a number of features related to different types of items and item

responses. As shown in Figure 4, approximately half of the JTs in our sample asked participants

to simply indicate whether each item was grammatical/acceptable or

ungrammatical/unacceptable. Some instruments (10%) also allowed participants to indicate if

they did not know whether an item was grammatical/acceptable or not. Another frequent

response elicited by JTs (29%) involves asking participants to respond on a scale such as -3

(completely unacceptable) to +3 (completely acceptable) (e.g., Yuan and Dugarova, 2012). In

addition, 23% of the JTs in our sample also instructed participants to identify the location of

errors when present, the majority of which (80%) also asked participants to attempt to correct the

errors they identified. Researchers utilized these data only sporadically and somewhat

idiosyncratically. For example, of the 87 instances in which corrections were made by

participants to incorrect items, some researchers analyzed incorrectly identified errors (21%),

some gave partial credit (9%), and others ignored them (2%). Most often, however, researchers

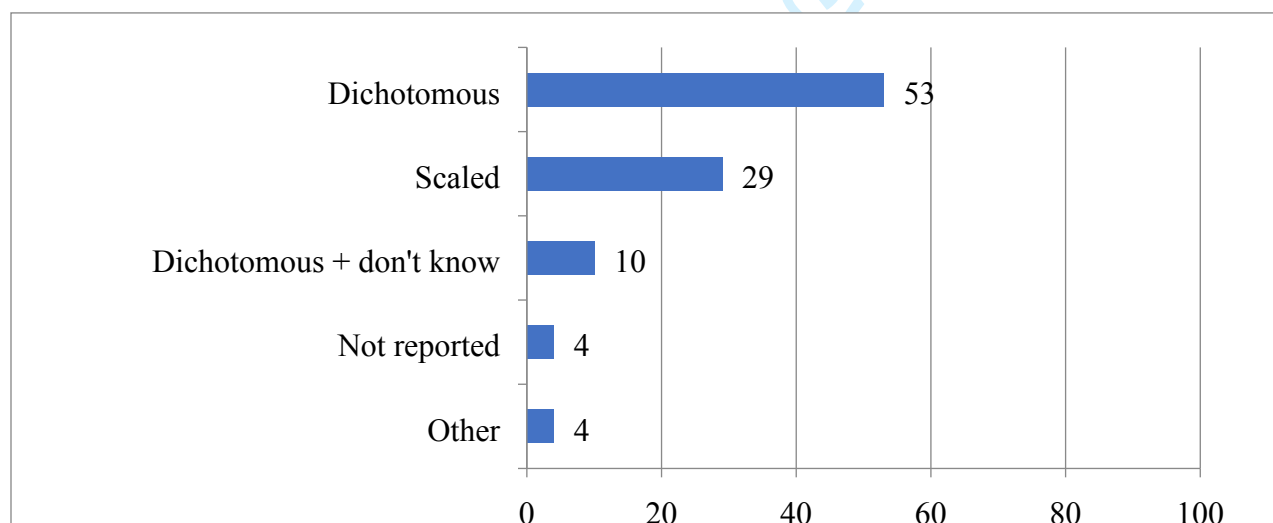simply did not report how such responses were handled (43%).

Figure 4. Percentage of different item response types

A small number of studies elicited additional responses following the main task of determining item acceptability or grammaticality. These include a rating of the confidence of their responses (5%), the basis for their response (4%, e.g., intuition/feel vs. rule), and providing a response to a meaning-based comprehension question (1%).

A number of the results thus far have examined the extent to which authors reported the details of their instrumentation. Also relevant to describing instrumentation is the reporting of reliability (i.e., internal consistency), which was found in only 16% of our sample. In most of these cases (81%), reliability was reported using Cronbach's alpha. In addition to coding studies for whether a reliability estimate was provided, we recorded each individual coefficient. The reliability estimates for internal consistency in the sample were then meta-analysed to examine their distribution. The bulk of the observed estimates fell in the .7-.9 range, with a median of .8 (see Figure 5). This value matches very closely Plonsky and Derrick's (2016) finding of .83 (interquartile range = .15) as the median estimate of reliability (internal consistency) for linguistic measures in L2 research.
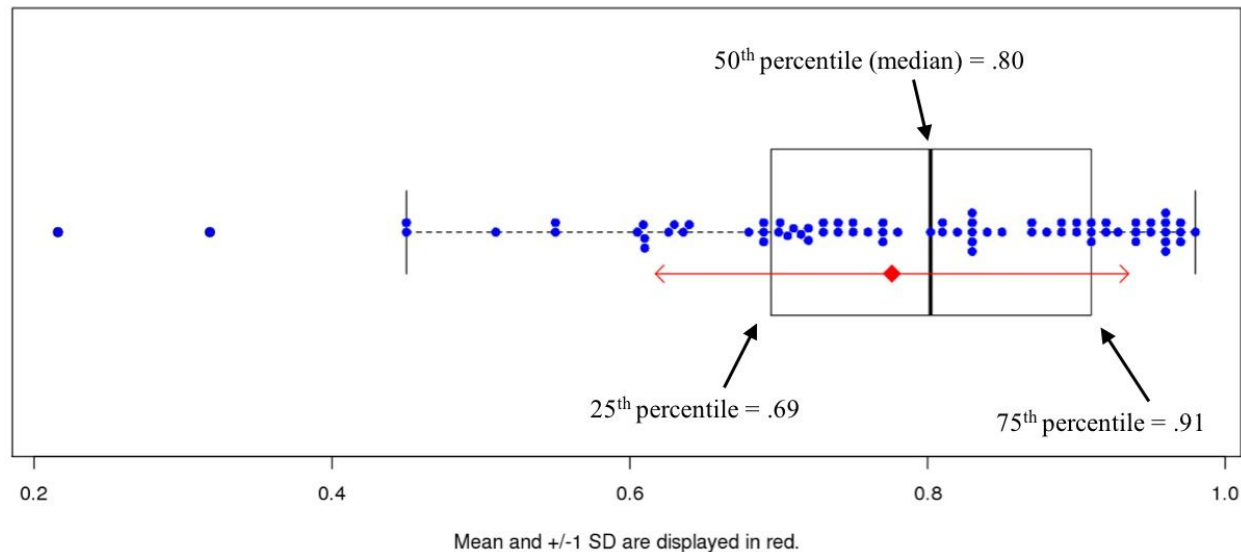
Figure 5. Reliability coefficients (internal consistency) from JTs in L2 research (k = 75)

56% also chose to administer additional measures of L2 learner knowledge. In most of

these cases, data from two (or more) instruments were analysed separately from the JT data. But

29% took a statistical approach to compare, correlate, or otherwise examine the relationships

among JTs and other measures of learner knowledge. For example, Spada, Shiu, and Tomita

(2015) ran bivariate correlations between acceptability JTs and a number of additional

instruments designed to measure L2 learners' implicit knowledge. Related to the approach in

Spada et al., we were interested in examining researchers' assumptions regarding the type(s) of

knowledge elicited by JTs. However, less than a quarter (24%) of the sample discussed this issue

overtly: 10% used JTs to measure knowledge labeled as implicit, procedural, or automatized;

nearly the same amount (9%) were interested in declarative or explicit knowledge; 4% of the

studies used JTs to measure knowledge in both categories. Ellis (2005), for example, sought to

measure learners' implicit and explicit knowledge using timed and untimed JT, respectively.

Discussion of implicit/explicit knowledge types were not evenly

distributed through our sample, with most occurring between

2010-16, as follows: 4 out of 12 studies from 1970-1989; 6 out of 54 from 1990-1999; 10 out

of 90 from 2000-2009; and 23 out of 142 from 2010-2016. The increase in the last decade might

arguably be attributable to the field's increased concern about explicit/implicit knowledge types

(Rebuschat and Williams 2012), but in fact the increase may be, more simply, due to the overall

increase in L2 studies using JTs, as the proportions have not increased.

As a final check on the methodological transparency of studies involving JTs, we also

coded for whether—and where—JTs were made available, thus facilitating inspection, analysis,

researcher training, and replication. Overall, 36% of the JTs were accessible, whether in primary

reports (22%), on IRIS (8%), and/or elsewhere such as an author's website (3%). Of course, the

flipside of this result is that 64% of the JTs in our sample were not available. Unfortunately,

there is no evidence of real improvement here over time. The percentage of JTs from our sample

that were available in some form are as follows: 1970s = 33%, 1980s = 57%, 1990s = 46%,

2000s = 30%, 2010s = 35%.

RQ3: To what extent do features of JTs (modality and timing) influence the scores they elicit?

The results thus far have indicated wide variability in terms of how JTs are designed and

implemented in L2 research. Of equal or greater importance is understanding whether variations

in the features of JTs impact learner performance. More concretely, we were interested in

ascertaining whether and to what extent different types of JTs might provide different

information about the amount of L2 knowledge learners demonstrate under different test

conditions. Very few studies have addressed this question. However, a meta-analytic approach

enables us to investigate relationships observed across studies, even if those relationships are not

the focus of any individual study. We present the results pertaining to two such features here: aural vs. written and timed vs. untimed conditions.

As shown in the results above, JTs are administered in the written modality much more frequently than aurally. In a small number of studies ($k = 17$), both modes have been employed. In those cases, a standardized mean difference (Cohen's d) was calculated between each sample's score on the two testing conditions, with the groups' pooled SD in the denominator when calculating d. Figure 6 presents the spread of d values, where a negative value indicates a higher score for aural JT scores and a positive value indicates a higher score for the same group's written JT. The median d value (.14) indicates that, overall, scores on written JTs tend to be slightly higher than for aural JTs. In addition, observed d values ranged from -.7 (a substantially higher score for an aural JT) to +1.28 (a large advantage for the written condition). The (unweighted) mean was .32 with a fairly a large SD (.59) and a 95% confidence interval (CI) [.02, .63] indicating little stability to the observed point estimate.
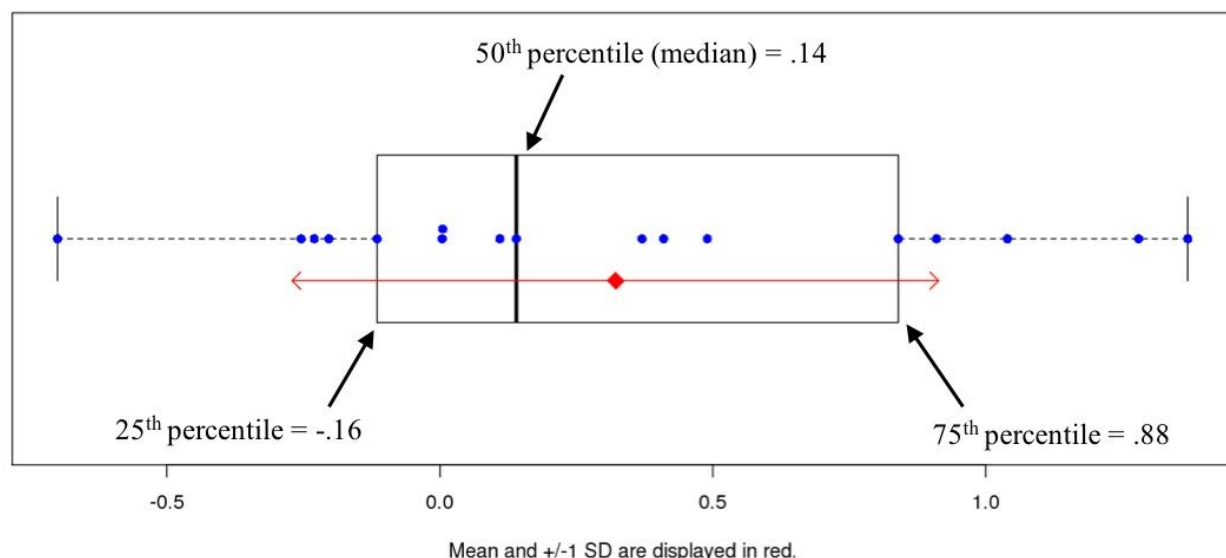
Figure 6. Within-group contrasts (d) of aural vs. written JT scores ($k = 17$; indicated in online Appendix with *)

The other feature we examine here with respect to its effect on JT scores was timing. Certain studies in our sample required participants to complete a judgment within a pre-determined period of time. This condition was often imposed—or not—in an attempt to tap a certain type of L2 knowledge (e.g., implicit or explicit; see Ellis, 2005). When results for both timed and untimed conditions were presented for the same participant samples, *d* values were calculated to index the difference in scores for each condition. Each of those *d* values appears in Figure 7. As observed for scores on aural vs. written JTs, the range of *d* values representing the difference between timed and untimed JTs was fairly wide. Unlike the comparison between modalities, however, all results were positive and showed a substantial discrepancy, with untimed JT scores tending to be much higher than timed JT scores. The (unweighted) mean was 1.6 (SD = 1.15), with a 95% CI of [1.09, 2.10] suggesting that this difference is fairly large and stable.

50th percentile (median) = 1.35

25th percentile = .64

75th percentile = 2.38

Mean and +/-1 SD are displayed in red.
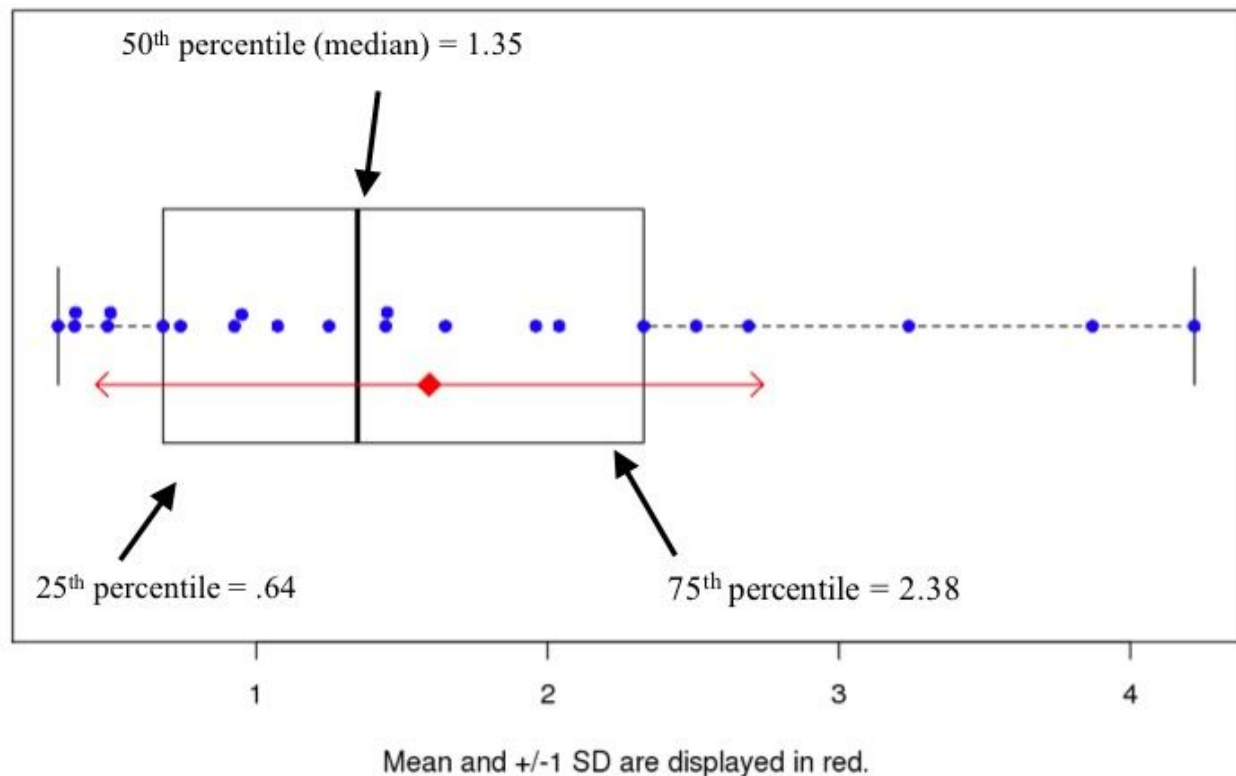
Figure 7. Within-group contrasts (*d*) of timed vs. untimed JT scores (*k* = 22; indicated in online Appendix with ^)

To address a question raised by a reviewer, we examined whether the effects of modality and time pressure might be conflated (e.g., whether aural JTs had been more likely to be timed). In total, only 22 JTs in the study sample were both timed and administered aurally. Of these 22 JTs, only three contributed to the comparison of possible modality effects (*d* = -.7, .14, 1.28) and only 1 contributed to the comparison of timing effects (*d* = .32). Generally, also, the 'written' versions in these modality comparisons were untimed. By definition, aural language happens at a fixed pace for all participants, so is, in a sense 'timed', and so, generally with these comparisons between aural and written, 'modality' is inevitably conflated with timing, unless one times them both to keep timing constant (as in Morgan-Short, Marsden, Heil et al. 2018), an issue raised

again in our Discussion. Note also that in the comparisons of timed vs. untimed, all the tests were written. Based on this follow-up analysis, we are fairly confident that such a conflation does not pose a threat to the validity of the results concerning modality- and timing-effects.

An additional step we took to further examine the validity of the effects found in our meta-analyses was to document the extent of the use of 'counterbalancing' to reduce confounds caused by administering one condition of the test before the other condition across all participants. For example, if the written modality was systematically administered after the aural modality, or the untimed modality after the timed modality, this could explain differences in the scores between the different conditions. To illustrate further, if the untimed JTs were always taken after the timed JTs and if those two tests were identical, practice effects could explain the large effects that we found above in favor of untimed JTs. Researchers can counterbalance using one or both of two approaches: one is at a within-test level (so the two conditions of the tests contain slightly different, though comparable, items, such as by keeping the grammar features constant but using different lexical items of a similar frequency); the other is at the test level (half the learners undertake one version first and the other of participants undertake the other condition first. In our subset of studies examining different timing conditions, the majority of studies (8/11 [73%]; 4 not reported) presented the timed JT before the untimed JT (many of which followed Ellis, 2005, procedurally). No studies reported any kind of counterbalance between timed and untimed presentation, and, also concerning, half used the same content for both. For the subset of studies examining different modalities, only two of six studies reported counterbalancing the presentation of aural and written stimuli, with written preceding aural twice and aural proceeding writing once (one not reported). Only two studies clearly indicated that they counterbalanced the items between the two modalities.

These findings reflect a fairly low level of reporting and/or use of counter-balancing in our samples of studies used in the meta-analyses. This systematicity between studies in administering one condition before the other is an important threat to the validity of the results of the meta-analysis. It should be noted that in terms of the comparison between timed and untimed JTs, the systematicity seen in presenting timed before untimed is heavily based on the impact of Ellis (2005), which a significant portion of our subset drew upon in developing their procedure. The potential impact of practice effects may, at least in part, explain (a) the lack of stable effects in the modality meta-analysis and (b) the strong effects found for untimed JTs producing higher scores than timed JTs.

Discussion

This study set out to describe the design and implementation of judgment tasks in L2 research. Our interest in doing so was motivated in part by their popularity in the field: JTs are employed frequently and increasingly in L2 research. Prior to the present study, however, the extent of their usage was unknown. Following a comprehensive search, we identified 302 published studies using 385 JTs. By comparison, Marsden et al. (2018b) identified 64 studies that have employed self-paced reading (SPR) tasks, Watanabe and Koyama's (2008) synthesis of cloze passage research was based on a sample of 33 studies; Yan, Maeda and Ginther (2016) surveyed 76 studies using elicited imitation and included 21 in their meta-analysis; and Bowles (2010) meta-analysed 14 samples using concurrent verbal reports (i.e., think-alouds).

The popularity of JTs in L2 research is likely due to a combination of at least three main factors, the first of which is versatility. Indeed, our results show that JTs have been used to examine learner knowledge on a variety of morphological and syntactic features. JTs have also

been applied frequently to support heavily debated claims regarding the nature of learner

knowledge (e.g., Vafaee et al., 2017). Such studies address long-standing, fundamental questions

posed and refined since the inception of the field (e.g., learning vs. acquisition; implicit vs.

explicit knowledge, declarative vs. automatized knowledge). However, our data showed that

only about a quarter of studies using JTs overtly discussed knowledge type. Further, although we

observed more of such discussion in more recent publications, we did not observe an increase in

terms of the *proportions* of JT studies that have alluded to knowledge type. To put this in

context, overt discussions of knowledge types was higher in studies reporting on self-paced

reading tests, where 26 out of 64 articles (41%) mentioned knowledge types, with 18 describing

the knowledge elicited as automatic/automaticity; 8 as implicit; and 14 as 'not explicit' (Marsden

et al., 2018b).

A second contributor to the popularity of JTs—and a consideration for any type of

language assessment (see Brown and Douglas, 2004)—is their practicality. JTs can target learner

knowledge on numerous features and can be administered and scored very quickly compared to

other elicitation methods. By contrast, much greater time and effort is required to elicit, code,

and score oral or written production, and with less guarantee of being able to elicit the target

feature or control the nature or amount of the feature elicited across participants. Further

contributing to their practicality as a research tool is the relative ease in comparing JT data across

participant groups (e.g., L1 vs. L2; low vs. high proficiency; learners who started learning at

different ages) relative, at least, to elicitations of L2 production.

Third, JTs appear to many to be easy to develop. It might be tempting to simply compose

a set of sentences that contain the target feature(s) and to then simply adjust some portion of

them to violate the structure(s) in question. As with other types of instruments that appear very

straightforward (e.g., gap fill, questionnaires), though, there are a number of considerations that

often go unattended and unexamined (see Spinner & Gass, in press). We have attempted in this

study to reveal and scrutinize many such considerations and features. In doing so we hope to

impress upon researchers using JTs that these choices merit not only consideration but

justification given the goals of the study, the substantive interests, the learner population (e.g.,

proficiency), assumptions about the type of knowledge being assessed, and so forth.

Regardless of the causes for its popularity, the widespread use of JTs requires that we

understand the extent of variation in their design and administration, and how such variation can

affect the data they elicit. The results of our study show that there has been very little consistency

in JT design and administration, even when similar linguistic features were being targeted. To

name a few of the design characteristics found to vary: item response types, item randomization,

the number of items per target feature, error identification, confidence ratings, comprehension

questions, task modality, and time pressure. In this respect, very little has changed since

Chaudron's early (1983) review of the use of JTs in L1 and L2 acquisition research, which

observed similarly disparate application of this instrument type. Studies in his sample exhibited,

for example, both binary and scaled judgements, timed vs. untimed conditions, and a wide

variety of target structures. In and of itself, this disparity is not a cause for concern, but our

review suggested that researchers are not sufficiently informed about whether and to what extent

such characteristics influence results.

To be clear, we are not arguing that all variants of JTs *necessarily* affect the data. We are

arguing, rather, that the lack of standardization across instruments presents a potential threat to

the validity of findings based on JTs and comparisons across studies. For example, our results for

RQ3 indicated that certain test characteristics do indeed influence results. All things being equal,

the results for RQ3 show that studies using untimed JTs can expect participant scores on average to be higher by one standard deviation than if they impose a time limit. To put this in more meaningful terms, the median difference between JT scores on timed and untimed conditions (*d* = 1.35)—when measuring the *same* target features from the *same* participants—is comparable to the difference typically observed between pre- and post-tests in L2 instruction research (Plonsky, 2017; Plonsky and Oswald, 2014). This result might also be interpreted as evidence of the validity of untimed and timed JTs as measures of more explicit and more implicit L2 knowledge, respectively. Making such a claim was not the purpose of our study. What is clear, though, is that our results do align with arguments that propose some discriminant validity of these two JT conditions.

On the other hand, our analysis of the influence on scores of another design feature, modality, did not find such a stark or reliable effect. Nevertheless, this finding is still useful for researchers making JT design decisions, as it shows that any a priori assumptions about the effect of modality require corroboration (see Johnson, 1992; Murphy, 1997). It is also illuminating in that it is, arguably, counter-intuitive. One might have expected aural JTs to impose more difficulties, all things being equal, as listening does not permit back-tracking/regression (other than that provided by a participant's own traces in their phonological store), and listening is often thought to provide less opportunity for participants to access controlled, explicit or metalinguistic knowledge (see discussion attention to form in oral vs. written mode in Morgan-Short, Marsden, Heil, et al. 2018).

Also critical to assessing validity, whether of a single instrument or a study, is the notion of methodological and psychometric transparency. In addition to being coded for different methodological variants, the primary studies in our sample were also coded for whether or not

those features were reported. The results were quite mixed. For example, 43% of the studies that asked participants to identify errors did not report how those data were then handled. The use of time pressure is another example and one that we know to influence task results. Not only did many studies fail to report whether JTs were timed or not, even among those that did report timing, few stated how the time limit was determined. Other frequent 'sins of omission' regarding task features include not reporting JT modality (18%) or format (i.e., computer-based or paper-and-pencil) (31%).

These results are not particularly surprising. Marsden et al. (2018b) observed a similarly spotty record for the reporting of features of self-paced reading tasks. For example, they found that sentence length was not reported in almost half the studies (30/64); 17/28 SPRs that used multi-word segments did not report the length or number of the segments; 25 of the 50 studies that used comprehension questions following critical trials provided no example of the questions. Similar to the current findings showing a lack of clarity or consistency about what was included in the JT analysis, Marsden et al. also found low levels of consistency (where consistency might be expected, i.e. between related studies) about which regions were analyzed and whether reaction times were measured for individual words or for the sum or means of multi-word segments. Further reflecting the current study's findings about the unequal or unreported balance between grammatical/acceptable and ungrammatical/unacceptable items, Marsden et al. found that in SPRs the balance between number of items in different conditions was not always reported, varied widely between studies, and did not align with recommended ratios. The extent to which conditions are balanced is important as too many items per condition can fatigue participants, desensitize them to ungrammaticality, or raise their awareness about the manipulations, whereas too few items per condition may fail to provide a stable result.

Another feature associated with transparency and instrument validity is reliability.

Unfortunately, Chaudron's (1983) blunt assessment that "reliabilities have generally not been

reported" (p. 368) still applies. Only 16% of the sample reported an estimate of instrument

reliability. This domain lags behind many others in L2 research (see Al-Hoorie and Vitta, 2018;

Plonsky and Derrick, 2016; Plonsky and Gass, 2011; but cf. Marsden et al., 2018b).

More important than the availability of reliability estimates is, of course, the actual

reliability of the instruments employed. In this context, reliability refers to the internal

consistency of the measurement tool. Higher reliability indicates greater consistency and

therefore less error in individual and group estimates of learner knowledge. In order to determine

the reliability across the JTs in our sample, we recorded and meta-analyzed each estimate that

was available. The median reliability coefficient was .8 (interquartile range = .22). This value

aligns very closely with Plonsky and Derrick's (2016) meta-analytic estimate of instrument

reliability throughout quantitative L2 research (.82, interquartile range = .15). We recognize,

though, that the reliability estimates that are available in published studies may not be

representative of the population of JT reliabilities; reported estimates are likely to be higher than

those that go unchecked and/or unreported.

A final issue related to methodological transparency, one closely related to addressing

some, if not all, of the 'sins of omission' discussed above, was the extent to which the full

instrument itself was available. We found 64% of our JTs were not available, thus restricting our

own capacity to synthesize some of the methodological decisions taken and the field's capacity

to scrutinize the validity and reliability of JT research. Although low, this is in fact a slightly

higher level of transparency than that found for SPR tests by Marsden et al. (2018b), where 77%

(49/64) of studies had only a brief example (e.g., one or two items) of stimuli available in the

article. This generally low level of transparency threatens the replicability of research, as replication studies either have to extrapolate full instruments from a few examples in the article, or re-construct the entire instrument, both routes introducing significant heterogeneity into any replication study (see Marsden et al., 2018a, for further discussion). This situation may change as more journals encourage their authors towards greater transparency (by, for example, adopting the Centre for Open Science badge scheme (Blohowiak et al., 2016) shown to increase long-term availability of materials and data (Kidwell et al., 2016)).

Also with a view to promoting greater transparency, and as a follow-up to the current study, we have established a 'special collection' of JTs on IRIS. The collection now holds 107 JTs from our synthesis and the metadata for 317 individual JTs (out of the total 385 JTs from 302 studies in our synthesis). Considering that many JTs were used for research published several decades ago, the generally positive response to our requests demonstrated to us a willingness of the field to engage in a more collaborative and synthetic effort. We hope that this JT collection will serve as a reference corpus for future syntheses and as resource for training and replication research.

Suggestions for Future Research

Our interest in this study was not solely in the use of JTs as matter of history. We are just as interested—if not more so—in future applications if this particular type of tool. With that interest in mind, we offer here a number of suggestions based on our results.

Our review has shown many of the same limitations observed elsewhere in L2 research in terms of target populations and demographics. For example, a small range of languages comprise the bulk of L1s and L2s involved (see Spinner, 2011). Somewhat unique to this body of research is the distribution of proficiency levels found in our sample. Whereas L2 research as a whole has

focused more often on relatively equal amounts of samples described as beginner/novice or

intermediate (Plonsky, 2017), the samples found in studies employing JTs were much more

likely to be labeled as advanced. One reason for this may be that there has tended to be interest in

advanced stages of the learning process in generative research, and 50% of our JTs were situated

within a generative framework (for example, generative-inspired research that focused on

whether L2ers can attain native-like levels of sensitivity to grammatical constraints).

Nevertheless, researchers employing JTs would do well to sample a wider range of

demographics and research contexts as a means to assess generalizability. We also suggest an

increase in the size of the samples in order to achieve more precise, reliable findings.

Our results also point to several suggestions pertaining to the design and implementation

of JTs. The general recommendation here is to strive for greater transparency with respect to the

features reported in our review. We would highlight among them the importance of piloting,

stating JT modality and format, stating whether the task was timed and, if so, how this was

calculated, making instruments available for future studies and replications, and calculating and

reporting instrument reliability. See Marsden et al. (2018a) for evidence that methodological

transparency makes research more replicable and renders findings from related studies more

interpretable. More fundamentally, we also call on our colleagues to make more deliberate and

justifiable choices regarding the design of their instruments, particularly when claims about

different types of L2 knowledge are involved. Reviewers and editors, too, must require greater

transparency in this regard. Study pre-registration provides yet another approach to avoiding

what often comes across as methodological idiosyncrasy. This step, taken before data collection

has begun, can demonstrate to readers that the design and analyses of a given study were

determined and registered a priori, and deviation from them did not occur as data was being

collected or analysed. For further explication on the benefits of and rationale for pre-registration and also, critically, the benefits of peer-review of studies before data is collected, we recommend Marsden, Morgan-Short, Trofimovich, and Ellis' (2018) editorial on registered reports in *Language Learning*. Outside of applied linguistics, we recommend Wagenmakers, Wetzels, Borsboom, van der Maas, and Klevit (2012) and Munafò et al. (2017).

The final area we recommend for future research also fits within the domain of the SLA-assessment interface (see Norris and Ortega, 2012; Gu, 2014). Recall from our results the relatively small number of studies (29%) that directly triangulated and/or compared the scores of JTs and other instruments intended to measure the same linguistic target(s) using one or more statistical analyses (see for example, Spada, Shiu, and Tomita, 2015). More research is needed here as well. Again, this recommendation is not new. Similar to Chaudron (1983), who argued that "judgments should be validated by other measures" (p. 369), we would also encourage authors to provide evidence of concurrent validity by administering multiple and complementary assessments of target constructs. Here, too, reviewers must play a critical role, insisting that researchers provide evidence of validity and reliability. We hope that the current climate of enhanced methodological awareness and assessment literacy (e.g., Grabowski and Dakin, 2014) will further encourage researchers in the field to embrace this idea.

Conclusion

Our study demonstrates a heavy reliance on acceptability judgement data in L2 research. Although our sample of studies reflected a wide variety of theoretical perspectives and linguistic features, JT usage seemed relatively unified in its purpose: to inform researchers about learners' sensitivity to norms in the input. However, our data have not allayed concerns about the wide variety of designs and administration conditions of JTs, as the effects of this variety on the

interpretability of JT data remain largely unknown. Partially addressing these concerns here, we

provided data on the likely effects of two key methodological decisions. We found that imposing

time constraints leads to substantially decreased scores, whereas modality did not strongly

influence outcomes. Regarding the other methodological characteristics that we examined (such

as the balance between grammatical/acceptable and ungrammatical/unacceptable conditions,

asking participants to locate or correct an unacceptable feature, providing scaled versus

dichotomous response options), we highlighted (a) the need for empirical scrutiny to build an

evidence base for more informed standardization in such methodological decisions and (b)

improved reporting and transparency of the methodological characteristics of JTs.

References

Al-Hoorie AH and Vitta JP (in press) The seven sins of L2 research: A review of 30 journals'

statistical quality and their CiteScore, SJR, SNIP, JCR Impact Factors. *Language

Teaching Research*.

Andringa S and Curcic M (2015) How explicit knowledge affects online L2 processing:

Evidence from differential object marking acquisition. *Studies in Second Language

Acquisition* 37(2): 237–268.

Bowles M (2010) *The Think-aloud Controversy in Second Language Research* (Chapter 3). New

York: Routledge.

Brown H and Douglas D (2004). Language Assessment: Principles and Classroom Practices.

New York: Longman.

Byrnes H (2013) Notes from the editor. *Modern Language Journal,* 97(4): 825-827.

Chaudron C (1983) Research on metalinguistic judgments: A review on theory, methods, and

results. *Language Learning* 33(3): 343-377.

Derrick D (2016) Instrument reporting practices in second language research. *TESOL Quarterly*

50(1): 132-153.

Ehrich JF and Meuter R I (2009) Acquiring an artificial logographic orthography: The beneficial

effects of a logographic L2 background and bilinguality. *Journal of Cross-Cultural*

Psychology 40(5): 711-745.

Ellis R (1991) Grammaticality judgments and second language acquisition. *Studies in Second*

Language Acquisition 13(2): 161–186.

Ellis R (2005) Measuring implicit and explicit knowledge of a second language: A psychometric

study. *Studies in Second Language Acquisition* 27(2): 141–172.

Gass S and Polio C (2014) Methodological Influences of *Interlanguage* (1972): Data then and

data now. In: Han Z-H and Tarone E (eds), *Interlanguage 40 Years later.* Amsterdam:

John Benjamins, pp. 147-171.

Gass S, Fleck C, Leder N, and Svetics I (1998) Ahistoricity revisited: Does SLA have a history?

Studies in Second Language Acquisition 20(3): 407-421.

Gu L (2014) At the interface between language testing and second language acquisition:

Language ability and context of learning. *Language Testing* 31(1): 111-133.

Godfroid A, Loewen S, Jung S, Park J-H, Gass S, and Ellis R (2015) Timed and untimed

grammaticality measure distinct types of knowledge: Evidence from eye-movement

patterns. *Studies in Second Language Acquisition* 37(2): 269-297.

Grabowski K and Dakin JW (2014) Test development literacy. In: Kunnan AJ (ed) *The*

Companion to Language Assessment. Oxford, UK: Wiley–Blackwell, pp. 751–768.

Gutiérrez X (2013) The construct validity of grammaticality judgment tests as measures of implicit and explicit knowledge. *Studies in Second Language Acquisition* 35(3): 423-449.

Hashemi M and Babaii E (2013) Mixed methods research: Toward new research designs in applied linguistics. *Modern Language Journal* 97(4): 828-852.

Johnson JS (1992) Critical period effects in second language acquisition: The effect of written versus auditory materials on the assessment of grammatical competence. *Language Learning* 42(2): 217-248.

Keating G and Jegerski J (2015) Experimental designs in sentence processing research. *Studies in Second Language Acquisition* 37(1): 1-32.

Kim J and Nam H (2017) Measures of implicit knowledge revisited: Processing modes, time pressure, and modality. *Studies in Second Language Acquisition* 39(3): 431-457.

Larson-Hall J (2017) Moving beyond the bar plot and the line graph to create informative and attractive graphics. *Modern Language Journal* 101(1) 240-270.

Larson-Hall J and Plonsky L (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning* 65(Supp. 1): 127-159.

Liu, Q and Brown D (2015) Methodological synthesis of research on the effectiveness of corrective feedback in L2 writing. *Journal of Second Language Writing 30*: 66-81.

Loewen S (2009) Grammaticality Judgment Tests and the Measurement of Implicit and Explicit L2 Knowledge. In: Ellis R, Loewen S, Elder, C, Erlam R, Philp J, and Reinders H (eds) *Implicit and Explicit Knowledge in Second Language Learning, Testing and Teaching*. Bristol, UK: Multilingual Matters, pp. 94–112.

Loewen S and Erlam R (2006) Corrective feedback in the chatroom: An experimental study, *Computer Assisted Language Learning* 19(1): 1-14.

Munafò, MR, Nosek BA, Bishop DV, Button KS, Chambers CD, du Sert NP, ... and Ioannidis JP (2017) A manifesto for reproducible science. *Nature Human Behaviour 1*: 0021.

Marsden E, Mackey A, and Plonsky L (2016) The IRIS repository: Advancing research practice and methodology. In: Mackey A and Marsden E (eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages* New York: Routledge, pp. 1-21.

Marsden E, Morgan-Short K, Thompson S, and Abugaber D (2018a) Replication in second language research: Narrative and systematic reviews, and recommendations for the field. *Language Learning* 68(2): 321–391.

Marsden E, Morgan-Short K, Trofimovich P, and Ellis NC (2018) Introducing registered reports at *Language Learning*: Promoting transparency, replication, and a synthetic ethic in the language sciences. *Language Learning* 68(2): 309-320.

Marsden, E, and Plonsky L (2018) Data, open science, and methodological reform in second language acquisition research. In: Gudmestad, A and Edmonds, A (eds.), *Critical Reflections on Data in Second Language Acquisition*. Philadelphia, PA: John Benjamins, pp. 219-228.

Marsden E, Thompson S, and Plonsky L (2018b) A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics* 39(5): 861-904

Marsden, E and Torgerson C (2012) Single group, pre- and post- research designs: Some methodological concerns. *Oxford Review of Education* 38(5): 583-616.

Morgan-Short K, Marsden E, Heil J, Issa, B., Leow, R. P., Mikhaylova, A., Mikołajczak, S.,

Moreno, N., Slabakova, R., & Szudarski, P. (2018) Attention to form and comprehension:

A multi-site replication. *Language Learning* 68(2): 392–437

Murphy VA (1997) The effect of modality on a grammaticality judgement task. *Second*

Language Research 13(1): 34-65.

Norris JM and Ortega L (2012) Assessing learner knowledge. In: Gass S M and Mackey A

(eds.), *Handbook of Second Language Acquisition.* New York: Routledge, pp. 573-589.

Norris, JM, Plonsky L, Ross SJ, and Schoonen R (2015) Guidelines for reporting quantitative methods

and results in primary research. *Language Learning* 65(2): 470-476.

Paquot M and Plonsky L (2017) Quantitative research mehods and study quality in learner

corpus research. *International Journal of Learner Corpus Research* 3(1): 61-94.

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.

Plonsky L (2014) Study quality in quantitative L2 research (1990-2010): A methodological

synthesis and call for reform. *Modern Language Journal* 98(1): 450-470.

Plonsky L (2017) Quantitative research methods. In: Loewen S and Sato M (eds.), *The Routledge*

Handbook of Instructed Second Language Acquisition. New York: Routledge, pp. 505-

521.

Plonsky L and Derrick DJ (2016) A meta-analysis of reliability coefficients in second language

research. *Modern Language Journal* 100(2): 538-553.

Plonsky L and Gass S (2011) Study quality in interactionist research. *Language Learning* 61(2):

325-366.

Plonsky L, and Ghanbar H (in press) Multiple Regression in L2 research: A methodological synthesis and

guide to interpreting R^2 values. *Modern Language Journal*.

Plonsky L and Gonulal T (2015) Methodological synthesis in quantitative L2 research: A review

of reviews and a case study of exploratory factor analysis. *Language Learning* 65(Supp.

1): 9-36.

Plonsky L and Oswald FL (2014). How big is 'big'? Interpreting effect sizes in L2 research.

Language Learning 64(4): 878-912.

Plonsky L and Oswald FL (2015) Meta-analyzing second language research. In: Plonsky L (ed.),

Advancing Quantitative Methods in Second Language Research. New York, NY:

Routledge, pp. 106-128.

Plonsky L and Oswald FL (2017). Multiple regression as a flexible alternative to ANOVA in L2

research. *Studies in Second Language Acquisition* 39(3): 579-592.

Rebuschat P and Williams J (2012) Implicit and explicit knowledge in second language

acquisition. *Applied Psycholinguistics* 33(4): 829-856.

Roberts L (2016) Self-paced reading and L2 grammatical processing. In: Mackey A and Marsden

E (eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for

Research into Second Languages*. New York: Routledge, pp. 58-73.

Rodriguez MC and Meada Y (2006) Meta-analysis of coefficient alpha. *Psychological Methods*

11(3): 306–322.

Schachter J, Tyson A, and Diffley F (1976) Learner intuitions of grammaticality. *Language

Learning* 26(1): 67–76.

Selinker L (1972) Interlanguage. *International Review of Applied Linguistics* 10(3): 209–231.

Shiu L-J, Yalçın, Ş, and Spada N. (2018). Exploring second language learners' grammaticality

judgment performance in relation to task design features. *System* 72: 215-225.

Spada N, Shiu JL-J, and Tomita Y (2015) Validating an elicited imitation task as a measure of

implicit knowledge: Comparisons with other validation studies. *Language Learning*

65(3): 723-751.

Spinner, P (2011) Second language research in Bantu languages: A (mostly) untapped research

opportunity. *Second Language Research* 27(3): 418-430.

Spinner P and Gass S (in press) Using *Judgments in Second Language Acquisition Research*.

New York: Routledge.

Sprouse J (2013) Acceptability judgments. *Oxford Bibliographies Online: Linguistics*.

http://www.oxfordbibliographies.com/view/document/obo-9780199772810/obo-

9780199772810-0097.xml

Vafaee P, Suzuki Y and Kachisnke I (2017) Validating grammaticality judgment tests: Evidence

from two new psycholinguistic measures. *Studies in Second Language Acquisition* 39(1):

59-95.

Wagenmakers EJ, Wetzels R, Borsboom D, van der Maas HL, and Kievit RA (2012) An agenda

for purely confirmatory research. *Perspectives on Psychological Science* 7(6), 632-638.

Watanabe Y and Koyama D (2008) A meta-analysis of second language cloze testing

research. *Second Language Studies* 26(2): 103-133.

Wheeler DL, Vassar M, Worley JA and Barnes LLB (2011) A reliability generalization meta-

analysis of coefficient alpha for the Maslach Burnout Inventory. *Educational and

Psychological Measurement* 71(1): 231–244.

Winke P (2014) Testing hypotheses about language learning using structural equation

modeling. *Annual Review of Applied Linguistics* 34: 102-122.

Yan X, Maeda Y, Lv J and Ginther A (2016) Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing* 33(4): 497-528.

Yuan B and Dugarova E (2012) Wh-topicalization at the syntax-discourse interface in English speakers' L2 Chinese grammars. *Studies in Second Language Acquisition* 34(4): 533–560.

Ziegler N (2016) Methodological practices in interaction in synchronous computer mediated communication: A synthetic approach. In: Mackey A and Marsden E (eds.), *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages*. New York: Routledge, pp. 197-223.

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Authors	Journal	Year	oral vs. written	timed vs. untimed
Abrahamsson	0	2012		
Abrahamsson & Hyltenstam	0	2007		
Abrahamsson & Hyltenstam	2	2009		
Abrahamsson & Hyltenstam	2	2009		
Akakura	6	2012		
Albirini & Benmamoun	18	2015		
Alibabaee, Momenzadeh, & Abarghoui	14	2014		
Al-Surmi	3	2012		
Ammar, Lightbown, & Spada	10	2010		
Anderson	53	2007		
Andringa, de Glopper, & Hacquebord	2	2011		
Andringa & Curcic	0	2015		
Ard & Gass	0	1987		
Athanasopoulos	33	2007		
Au, Oh, Khightly, Jun, & Romo	39	2008		
Balcom	16	1997		
Balcom	61	1999		
Balcom	61	1999		
Bardovi-Harlig & Dornyei	13	1998		
Barrios & Bernaro	40	2012		
Batterink & Neville	22	2013		
Bautista	25	2004		
Bialystok	2	1979		
Bialystok	16	1997		
Bialystok & Majumder	17	1998		
Bialystok & Miller	18	1999	*	
Bialystok & Miller	18	1999	*	
Bianchi	18	2013		
Birdsong	38	1992		
Bley-Vroman, Felix, & Ioup	16	1988		
Bley-Vroman & Yoshinaga	16	2000		
Bordag & Pechmann	18	2007		
Borgonovo, Bruhn de Garavito, &	0	2015		
Bowles	0	2011		^
Bowles	0	2011		^
Bruhn de Garavito & Valenzuela	18	2008		
Bylund, Abrahamsson, & Hyltenstam	0	2012		
Cho	27	2016		
Cho & Slabakova	1	2015		
Choi & Clark	19	2006		
Chung	16	2012		
Clahsen, Martzoukou, & Stavrakaki	16	2010	*	
Clahsen, Martzoukou, & Stavrakaki	16	2010		
Clahsen et al.	16	2013		
Conradie	16	2006		

Conroy & Cupples	0	2010	
Cook	4	2003	
Cormier et al.	55	2012	
Coughlin & Tremblay	17	2013	
Cox	1	2015	
Cox & Sanz	4	2015	
Cuza	27	2012	
Cuza & Frank	56	2011	
Cuza & Frank	16	2015	
Cuza et al	27	2012	
Cuza, Perez-Leroux, & Sanchez	0	2012	
Davis & Kaplan	1	1998	
deGraff	0	1997	
deGraff	0	1997	
DeJong	0	2005	
DeKeyser	0	2000	
Dekeyser, Alfi-Shabtay, & Ravid	17	2010	
Dekeyser, Alfi-Shabtay, & Ravid	17	2010	
Dekydspotter, Sprouse, & Anderson	53	1999	
Della Putta	0	2016	
Derwing, Rossiter, & Ehrensberger-	10	2002	
Derwing, Rossiter, & Ehrensberger-	10	2002	
Doughty	0	1991	
Doughty	0	1991	
Duffield & Matsuo	0	2009	
Duffield & White	16	1999	
Dugarova	16	2014	
Dussias & Pinar	16	2010	
Ehrich & Meuter	50	2009	
Ellis	0	1991	
Ellis	0	1991	
Ellis	0	2005	
Ellis	0	2005	^
Ellis	1	2006	^
Ellis	1	2006	
Ellis, Loewen, & Erlam	0	2006	
Erçertin & Alptekin	17	2013	^
Erçertin & Alptekin	17	2013	
Erlam & Loewen	20	2010	
Falk & Bardel	16	2010	
Fallah, Jabbari, & Fazilatfar	16	2015	
Felsner, Roberts, Marinis, & Gross	17	2003	
Felser, Sato, & Bertenshaw	18	2009	^
Felser, Sato, & Bertenshaw	18	2009	
Finney	16	1997	
Flege & Liu	0	2001	

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Flege, Yen-Komshian, & Liu	39	1999
Fotos	13	1994
Fotos & Ellis	13	1991
Gabriele	0	2009
Gabriele	0	2009
Gabriele & Canales	28	2011
Gao	21	2014
Garcia-Mayo & Slabakova	27	2014
Gass	2	1979
Gass	13	1983
Giancasparo, Halloran, & Iverson	18	2015
Godfroid et al.	0	2015	^
Godfroid et al.	0	2015	^
Goo	0	2012
Gor et al	65	2007
Granena	1	2014	^
Granena	1	2014
Grey, Cox, Serafini, & Sanz	8	2015
Grey, Williams, & Rebuschat	0	2014
Grey, Williams, & Rebuschat	0	2014
Guijarro-Fuentes	18	2012
Guijarro-Fuentes & Larranaga	27	2011
Guion & Flege	17	2000
Gutierrez	61	2012	^
Gutierrez	61	2012	^
Han & Ellis	6	1998	^
Han & Ellis	6	1998
Hara	16	2007
Hawkins	16	1987
Hawkins & Chan	16	1997
Hawkings, Towel, & Bazergui	16	1993
Hedgcock & Lefkowitz	30	1993
Hermas	63	2010
Hermas	63	2010
Hermas	63	2010
Hermas	63	2010
Hermas	63	2014
Hermas	63	2014
Hermas	63	2014
Hermas	63	2014
Hermas	63	2014
Hermas	63	2014
Hermas	27	2015
Hermas	27	2015
Hopp	16	2005
Hopp	16	2006

Hopp	18	2009
Hopp	28	2010
Hopp	28	2010
Hu & Liu	16	2007
Huang	34	2014
Hulk	16	1991
Hwang & Lardiere	16	2013
Hwang & Lardiere	16	2013
Hylstenstam, Bylund, Abrahamsson, &	18	2009
Hylstenstam, Bylund, Abrahamsson, &	18	2009
Hylstenstam, Bylund, Abrahamsson, &	18	2009
Indrarathne & Kormos	0	2016
Inagaki	0	2001
Inagaki	0	2001
Ionin & Montrul	2	2010
Ionin & Wexler	16	2002
Ioup et al.	0	1994
Isabelli-Garcia	30	2010
Ivanov	16	2012
Izumi	2	2003
Izumi, Bigelow, Fujiwara, & Fearnow	0	2010
Izumi & Lakshmanan	16	1998
Jabbari & Salimi	48	2015
Jackson & Roberts	17	2010
Jia & Aaronson	17	2003
Jia, Aaronson, & Wu	17	2002	*
Jia, Aaronson, & Wu	17	2002
Johnson	2	1992
Johnson & Newport	66	1989
Johnson & Newport	55	1991
Johnson & Newport	55	1991
Johnson, Shenkman, Newport, & Medin	39	1996
Juffs	16	1998
Juffs	68	2004
Juffs	16	2005
Juffs & Harrington	2	1996
Jung	6	2016
Kadia	13	1988
Kadia	13	1988
Kang	4	2009
Khatib & Nikouee	32	2012
Kovelman, Shalinksy, Berens, & Petitto	35	2008
Kovelman, Shalinksy, Berens, & Petitto	35	2008
Kupferborg & Olshtain	10	1996
Kupisch	18	2012
Kweon & Bley-Vroman	16	2011

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Lan, Sung, Cheng, & Chang	7	2015
Larson-Hall	16	2008
Leon, Tsung, & Tse, Shum, & Ki	27	2012
Leon, Tsung, & Tse, Shum, & Ki	27	2012
Leung	18	2005
Leung	18	2005
Li	8	2013
Li, Ellis, & Zhu	29	2016
Lightbown & Spada	10	2000
Lightbown & Spada	10	2000
Loewen & Erlam	9	2006
Loewen & Erlam	9	2006
Lozano	69	2002
Lozano	69	2002
Lozano	16	2006
Mai & Yuan	16	2016
Mai & Yuan	16	2016
Mandell	16	1999
Marian, Blumenfeld, & Kaushanskaya	37	2007
Marsden	16	2008
Marsden	58	2009
Marsden & Chen	2	2011
Masny	10	1997
Masny & d'Anglejan	34	1985
Masny & d'Anglejan	34	1985
Mayberry & Lock	47	2003
McDonald	17	2000
McDonald	39	2006
McDonald & Roussel	18	2010
McManus	4	2013
McManus & Marsden	0	2016
McManus & Marsden	0	2016
Meunier & Littre	8	2013
Meunier & Littre	8	2013
Monteiro	7	2014
Monteiro	7	2014
Montrul	16	1999
Montrul	27	2006
Montrul	27	2006
Montrul	17	2010
Montrul	16	2010
Montrul	18	2014
Montrul & Bowles	18	2009
Montrul & Bowles	18	2009
Montrul & Bowles	56	2010
Montrul et al.	18	2014

Montrul & Ionin	18	2010	
Montrul & Ionin	8	2012	
Montrul & Ionin	8	2012	
Morgan et al.	23	2002	
Morgan-Short et al.	18	2014	
Murphy	16	1997	*
Murphy	16	1997	*
Murphy	16	1997	
Murphy	16	1997	
Nabei & Swain	10	2002	
Nagano	59	2015	
Nagano	59	2015	
Neubauer & Clahsen	0	2009	
Oh	16	2010	
Omaki & Schulz	0	2011	
Paap & Liu	43	2014	
Pae et al.	34	2013	
Papadopoulou & Clahsen	0	2003	
Papadopoulou & Clahsen	0	2003	
Papadopolou et al.	16	2010	
Paradis	2	2010	
Paribakht & Bingham Wesche	54	1993	
Park & Starr	63	2015	
Parodi & Tsimpli	16	2005	
Parodi & Tsimpli	16	2005	
Parodi & Tsimpli	16	2005	
Perpiñán	18	2015	
Pfenniger	62	2014	
Pires & Rothman	27	2009	
Polomska	16	1988	
Potowski, Jegerski, & Morgan-Short	2	2009	
Prentza	53	2014	
Rassaei	8	2014	
Rebuschat & Williams	17	2012	
Rebuschat & Williams	17	2012	
Reinders	6	2009	
Renou	30	2000	*
Renou	30	2000	*
Renou	10	2001	*
Renou	10	2001	*
Revesz	2	2012	
Rinke & Flores	18	2014	
Robenalt & Goldberg	2	2016	
Roberts & Felsner	17	2011	
Roberts, Gullberg, & Indefrey	0	2008	
Roberts & Liszka	16	2013	

Robinson	0	1996	
Robinson	0	1997	
Robinson	0	2005	
Robinson	0	2005	
Robinson & Ha	0	1993	
Robinson & Ha	0	1993	
Rogers	13	2015	
Rogers, Revesz, & Rebuschat	17	2016	
Rothman	27	2007	
Rothman	27	2009	
Rothman	27	2009	
Rothman & Cabrelli Amaro	16	2010	
Rothman & Cabrelli Amaro	16	2010	
Rothman & Cabrelli Amaro	16	2010	
Rothman & Iverson	64	2007	
Rothman & Iverson	0	2013	
Rothman & Iverson	0	2013	
Sabourin, Stowe, & de Haan	16	2006	
Sagarra & Herschensohn	28	2010	
Sagarra & Herschensohn	2	2011	
Sajjadi & Tahririan	4	2001	
Salaberry	4	2013	
Santoro	16	2007	
Sanz et al	2	2009	
Schachter	16	1976	
Schachter	16	1990	
Schachter & Yip	0	1990	
Schulz	16	2011	
Schulz	16	2011	
Schmid, Gilbers, & Nota	16	2014	
Shirai & Kuruno	2	1998	
Slabakova	16	1999	
Slabakova	16	2003	
Slabakova	16	2006	
Slabakova	16	2006	
Slavoff & Johnson	0	1995	
Snape	18	2008	
Soler	59	2014	
Soler	27	2015	
Soler	27	2015	
Sopata	18	2005	
Spada, Shiu, & Tomita	2	2015	*
Spada, Shiu, & Tomita	2	2015	*
Spinner	2	2013	
Spinner	2	2013	
Stafford	1	2013	

Stafford, Bowden, & Sanz	2	2011	
Stringer et al.	16	2011	
Svalberg	25	1998	
Tagarelli	0	2016	
Tanner, McLaughlin, Herschensohn, &	18	2013	
Tolentino & Tokowicz	2	2014	
Tolentino & Tokowicz	2	2014	
Torito	18	2001	
Toth	0	2000	
Toth	2	2006	
Toth	2	2008	
Toth & Guijarro-Fuentes	17	2013	
Towell, Hawkins, & Bazergui	0	1993	
Trahey	16	1996	
Trahey & White	0	1993	
Tremblay	16	2006	
Tsang	27	2015	
Uziel	16	1993	
Vafaee, Suzuki, & Kachisnke	0	2016	^
Vafaee, Suzuki, & Kachisnke	0	2016	^
Vulchonova, Talcott, Vulchanov, &	43	2012	
Westergard et al.	27	2016	
White	2	1985	
White	0	1987	
White	0	1987	
White	0	1987	
White	58	1990	
White	58	1990	
White et al.	2	1997	
White et al.	2	1997	
White et al.	2	1997	
White et al.	59	2012	
White & Genesse	16	1996	
White & Ranta	10	2002	
Whong-Barr & Shwartz	0	2002	
Winitz	8	1996	
Xu	16	2014	
Yalcin & Spada	0	2016	
Yalcin & Spada	0	2016	
Yang & Givon	0	1997	
Yang & Li	45	2012	
Yeganeh, Ghoreyshi, & Darabi	31	2013	
Yen-Konshian, Robbins, & Flege	17	2001	
Yuan	2	1995	
Yuan	18	2007	
Yuan	16	2007	

Yuan	16	2010
Yuan	27	2013
Yuan	16	2014
Yuan & Dugarova	0	2012
Yuan & Zhao	27	2010
Zhang	8	2012
Zhang	21	2015
Zhang	21	2015
Zhang	0	2015
Zhang	0	2015
Zhang	17	2015
Zhao	16	2014
Zhao	16	2014
Zyzik	16	2008

Appendix A

List of 69 Coded Features with Reliability Estimates

Publication Information

1. *Authors*

2. *E-mail*

3. *Journal*

4. *Year of publication*

Study Context

5. *context* (κ = .757; 92% agreement) – are participants learning the target language for a class or just for the study? (0 = classroom learner, 1 = learned for study, 2 = non-classroom learners of the language)

6. *shot_dev* (κ = .946; 95% agreement) – are participants tested for knowledge at a specific moment in time or to measure development over time (0 = one shot on "static Knowledge", 1 = developmental)

7. *effect* (83% agreement, too few observations to determine κ) – If developmental, was the GJT used to measure effectiveness of intervention/treatment? (0=not used as an outcome measure, 1=yes used as an outcome measure)

8. *delay_ptest* (100% agreement) – If used as outcome measure, was a delayed post-test conducted? (0 = no, 1 = yes)

9. *comp_prof* (κ = .838) (94% agreement) – Was data collected and compared across proficiency levels? (0 = no, 1 = yes)

10. *L1* (κ = .940) (95% agreement) – What was the first language of participants?

1

11. *L1s_Comp* (limited observations) – If multiple L1s amongst the learners (between-

   subject), were their results compared (i.e., differences across L1 groups)? (0 = no, 1 =

   yes)

12. *L2_if_L3* (limited observations) – If target language was a third language, what was the

   second language of participants?

13. *TL* (κ = .862) (89% agreement) – What was the target language of the study?

14. *SL_FL* (κ = .738) (84% agreement) – What was the learning context of the study? (0 =

   SL context, 1 = FL context, 2 = not real, a priori learners (e.g. a new, real language but

   first exposure), 3 = artificial language, 999 = not reported)

15. *htg* (κ = .849, 98% agreement) – Were the participant heritage language learners (0 = no,

   1 = yes)

16. *beg* – (κ = .423, 76% agreement) Were participants at beginner proficiency as reported by

   author (or up to 2 years of study) (0 = no, 1 = yes, 2 = not reported)

17. *int* – (κ = .671, 81% agreement) Were participants at intermediate proficiency as reported

   by author (0 = no, 1 = yes, 2 = not reported)

18. *adv* – (κ = .639, 79% agreement) Were participants at advanced proficiency as reported

   by author (or up to 2 years of study) (0 = no, 1 = yes, 2 = not reported)

19. *unint* (κ = .632, 94% agreement) – Was the proficiency level uninterpretable from the

   description provided? (0 = no, 1 = yes)

20. *profJust* (κ = .651, 79% agreement) – How did the authors justify their proficiency

   labels? (0=no justification (opinion), 1=assumed from educational/intitutional level,

   2=standardised proficiency test (IELTS, TOEFL, ACTFL, A-level, CEFR, etc.), 3=

   other)

21. *age* (κ = .774, 94% agreement) – What was the age range of participants? (0 = child: 0-1,

    1 = teen: 13-17, 2 = adult 18+, 3 = multiple, 4 = not reported)

22. *inst* (κ = .811, 90% agreement) – What was the institutional level of participants? (0 =

    elementary/primary (up to 13'ish), 1 = secondary (14-18'ish), 2 = tertiary

    college/university, 3 = language institute, 4 = not classroom learners/users, 999 = not

    reported)

23. *NS_ctrl* (κ = .822, 92% agreement) – Were NNS GJTs compared to a NS "control"

    group? (0 = no, 1 = yes)

24. *N_L2* (Cronbach's Alpha = .995) – What was the sample size of the L2 participants?

25. *N_NS* (Cronbach's Alpha = .960) – What was the sample size of the native speaker

    control group?

Judgment Task Design

26. *knowledge* (κ = .622, 87% agreement) – Were assumption made about type(s) of

    knowledge being tapped? (0 = no, 1 = yes, implicit, procedural, and/or automatized, 2 =

    yes, explicit and/or declarative, 3 = yes, 1+2, 4 = yes, but not clear whether 1 or 2)

27. *K_targeted* (Cronbach's Alpha = .390) – Number of items in judgment task not including

    distractors

28. *K-ungramm* (Cronbach's Alpha = .946) – Number of K-targeted items that were

    ungrammatical

29. *K_distractors* (Cronbach's Alpha = .656) – Number of items in judgment task not

    including distractors

30. *break* (κ = .571, 94% agreement) – Were participants offered breaks during the task? (0 =

    no, 1 = yes, 2 = not reported)

3

31. *mode* (κ = .686, 84% agreement) – What modality was the judgment task delivered in? (0

= aural, 1 = written, 2 = both, 3 = not reported, 4 = visual, 5 = half aural/half written)

32. *format* (κ = .595, 71% agreement) – How were item responses recorded? (0 =

paper/pencil, 1 = computer-based, 2 = unreported, 3 = spoken)

33. *#_of_features* (Cronbach's Alpha = .798) – How many grammatical features were

targeted in the judgment task?

34. *morphosyn_type* (κ = .780, 81% agreement) – What were the morphosyntactical target(s)

of the judgment task?

35. *incidental* (κ = .625, 85% agreement) – Was the morphosyntactic target incidental to the

overall goal of the study? (0 = No, 1 = Yes)

36. *timelimit* (κ = .647, 79% agreement) – Was a limit given for how long the participant had

to complete an item? (0 = no, 1 = yes, 2 = not reported)

37. *time_calc* (κ = .449, 67% agreement) – If time limit used, was time calculated based on

(0 = native speaker norms, 1 = non-native norms, 2 = other, 3 = not reported)

38. *adjust_time* (100% agreement) – If native norms used for timing, what percentage of time

to complete a judgment was added? (0 = not reported, 1 = 10%, 2 = 20%, 3 = 30%. 4 =

1.2 x NS RT)

39. *time_provided* – (100% agreement) For Timed JTs, the length of time provided to

respond to an individual item (in seconds).

40. *timed* (κ = .438, 68% agreement) – Was how long they took to complete an item

measured? (0 = no, 1 = yes, 2 = not reported)

41. *identify* (κ = .294, 54% agreement) – Were participants asked to identify in some way

incorrect items? (0 = no, 1 = yes, 2 = not reported)

42. *fix* (κ = .294, 54% agreement) – If asked to identify incorrect statements: do participants…? (0 = nothing, 1 = fix the error, 2 = just indicate problem area (e.g., circle it), 3 = both)

43. *timing* – (limited observations) If corrections were made to incorrect items, when were they made? (0 = none/not relevant, 1 = corrections made WHILE making judgments, 2 = corrections made afterwards, 3 = both)

44. *InCorrCorrection* – (limited observations) If corrections were made, what did researchers do with INCORRECT corrections? (0 = not reported, 1 = marked correct, 2 = marked incorrect, 3 = removed from analyses, 4 = errors analyzed, 5 = Partial Scoring Used)

45. *CB_forms* (κ = .109, 50% agreement) – Were 2 counter-balanced but otherwise identical forms of the JT used? (0 = no, 1 = yes, 2 = not reported)

46. *item_rand* (κ = .399, 66% agreement) – Were Items randomized across participants? (0 = no, 1 = yes, 999 = not reported)

47. *bal_corr* (κ = .614, 74% agreement) – Were the number of correct and incorrect items balanced? (0 = equal, 1 = unequal, 2 = not reported)

48. *context*? (κ = .796, 95% agreement) – Were items presented within a larger context? (0 = no context reported, 1 = paired with picture/video, 2 = contextual sentence/story provided, 3 = targets embedded within larger prose [e.g., passage correction])

49. *response_type* (κ = .759, 85% agreement) – What response type was utilized? (0 = dichotomous [correct vs. incorrect, [un]acceptable], 1 = scale of acceptability/grammaticality, 2 = dichotomous + "I don't know"/not sure option, 3 = not reported, 4 = other, see scale descriptor)

50. *scale* (limited observations) – If a scale of acceptability was used, how many options

were on the scale (e.g., 5)?

51. *scale_analysis* (limited observations) – Was the scale analyzed as described, or as

dichotomous? (1 = as described, 2 = dichotomous, 3 = mean score, 4 = other, 999 = not

reported)

52. *jdg_con* (κ = 1.000, 100% agreement) – Did participants rate confidence of response? (0

= no, 1 = yes)

53. *jdg_basis* (κ = 1.000, 100% agreement) – Did participants rate basis of judgment (e.g.,

intuition/feel vs. rule)? (0 = no, 1 = yes)

54. *compQ* (κ = 1.000, 100% agreement) – Were participants asked to answer comprehension

question after judgment? (0 = no, 1 = yes)

55. *compQ_ornt* – (limited observations) What was the learners' attention oriented to in the

comprehension question? (0 = meaning, 1 = more info in addition to meaning, 2 = other,

3 = not reported)

56. *instructions* – (open response) What were the instructions for the JT?

57. *instruct_source* (κ = .381, 89% agreement) – List the instructional source for

Instructions. (0 = paraphrase, 1 = description from text, 2 = exact directions)

58. *gram_ungramcoding* (κ = .65, 82% agreement) – Are grammatical and ungramamtical

items analyzed separately? (0 = no, 1 = yes)

59. *rel* (κ = .896, 97% agreement) – Was a reliability estimate for the judgment task

provided? (0 = no, 1 = yes)

60. *rel_index* – (100% agreement) If reliability reported, which index was used? (0 = none, 1

= Cronbach's alpha, 2 = interrater (%), 3 = other, 4 = both, 999 = Not reported)

61. *rel_est* – (100% agreement) If reported, what was the reliability estimate given?

62. *Other_instr* (κ = .740, 89% agreement) – Was another instrument used to triangulate data

from judgment task? (0 = no, 1 = yes)

63. *other_inst_type* – (open response) If other instrument, what kind of OTHER instrument

was used?

64. *other_inst_comp* (κ = .721, 84% agreement) – If other instrument used, did authors

attempt to correlate/compare with other instrument(s) in a statistical way? (0 = no [just

part of a battery], 1 = yes [some statistical correlation, validation, comparison made])

65. *task_avail* (κ = .732, 89% agreement) – Is the judgment task publically available? (0 =

no, 1 = yes, in the article, 2 = yes, on IRIS, 3 = yes, in the article + IRIS, 4 = yes,

somewhere else [e.g., authors' website])

66. *multi_JTs* (κ = .453, 74% agreement) – Did the study compare results of multiple

judgment tasks to test different task features/conditions? (0 = no, 1 = yes)

67. *name_of_test* – (open response) What terminology did the authors use to refer to their

judgment task?

68. *test_code* (κ = .566, 78% agreement) – What terminology did the authors use to refer to

their judgment task? (0 = No Title Provided, 1 = Grammaticality Judgement Test/Task, 2

= Acceptability Judgement Test/Task, 3 = Truth-Value Judgement Test, 4 = Forced-

Choice Elimination Task, 5 = Grammar Preference Task, 6 = Aural Grammar Test, 7 =

Questionnaire, 8 = Grammar Scan, 9 = Sentence Judgement/Interpretation Task, 10 =

Aural Judgement Task, 11 = Written Judgement Task, 12 = Sentence Completion Task,

13 = Moving Window Experiment, 14 = Story (Compatibility) Task, 15 = Aural Priming

Task, 16 = Context Evaluation/Matching Task, 17 = Forced Choice Task, 18 =

Explanation Task, 19 = Correction Task, 20 = Working Memory Reading Span Task, 21

= Picture Sentence Acceptability/Judgement) Task, 22 = Behavioral ERP Assessment, 23

= Plausability Rating Task, 24 = Picture Description Task, 25 = Passage Correction Task,

26 = Word Order Judgement Task, 27 = Acceptabilty Rating Task, 28 = Context

Felicitousness Task, 29 = Preference Task, 30 = Picture Matching/Selection Task, 31 =

Self-Paced Reading Task)

69. *framework* (κ = .757, 77% agreement) – What theoretical framework did the authors

employ? (0 = Unclear, 1 = Formal, 2 = Functional, 3 = Usage-Based, 4 = Skill-

Acquisition, 5 = Input Processing and Processing Instruction, 6 = Declarative/Procedural,

7 = Processability, 8 = Interactionist, 9 = Sociocultural, 10 = Complexity, 11 =

Implicit/Explicit, 12 = Working Memory, 13 = Neurolinguistic Processing, 14 =

Connectionist, 15 = Neurocognitive Disorder)