CrossMark

Annual Review of Applied Linguistics (2014), **34**, 37–64.
© Cambridge University Press, 2014, 0267-1905/14 $16.00
doi: 10.1017/S0267190514000026

Judgment and Interpretation Tasks in Second Language Research

Tania Ionin and Eve Zyzik

This article provides an overview of recent studies in second language acquisition that use tasks that elicit learners' judgments about the grammaticality of language or learners' interpretation of language. We discuss acceptability judgment tasks, preference tasks, truth-value judgment tasks, and other types of interpretation tasks. For each task type, recent studies that use that task are briefly summarized, with a focus on advantages and disadvantages of the methodology in relation to the study's objectives. A variety of topics related to task administration are covered, including (but not limited to) different types of rating scales; presentation of target sentences in isolation versus in the context of other sentences, stories, and/or pictures; visual versus auditory modality of presentation; and timed versus untimed tasks.

The most direct information about language learners' linguistic knowledge is, arguably, production: if learners produce a particular linguistic expression (such as a word, a phrase, or a sentence), it is reasonable to conclude that this expression is part of the learners' linguistic knowledge. Indeed, the study of learners' production has a long-standing tradition in second language acquisition (SLA), going back at least as far as the morpheme order studies of the 1970s, as well as more recent work (see Mackey & Gass, 2005 for an overview). In fact, Selinker (1972) argued that analyses of interlanguage should be based on observable output, that is, learners' attempted production in "meaningful performance situations" (p. 214). At the same time, production data have many limitations. The fact that a learner has not produced a particular expression, or has made a production error, does not necessarily reflect a lack of linguistic knowledge: Other factors, such as avoidance, phonological complexity, or difficulty with retrieval from memory (to name but a few) may be responsible. Conversely, the production of certain frequent formulaic expressions may reflect rote memorization rather than linguistic knowledge. For these reasons (among others), SLA researchers may use other data collection tools to study learners' linguistic knowledge. This article focuses on two main types of tasks that do not involve production: tasks that elicit judgments of grammaticality or acceptability and

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

tasks that target interpretation. Other tasks that fall beyond the scope of this article include measures of elicited production (which cover a wide range, from fill-in-the-blank tests to elicited imitation tasks) as well as psycholinguistic tools (such as self-paced reading and eye-tracking).

The terms grammaticality judgment task (GJT) and acceptability judgment task (AJT) are often used interchangeably in the literature to mean a task in which participants are presented with one sentence at a time and asked to judge its grammaticality. However, Cowart (1997) argued that the term AJT is more appropriate than the term GJT. According to Cowart, grammaticality is an abstract concept (a sentence either is or is not grammatical in a given grammar), which cannot be tested directly. Experimental tasks can test the acceptability of a given sentence by speakers, and the speakers' judgments can consequently be used to make inferences about grammaticality. Following Cowart, we use the term AJT rather than GJT in this article, but we note that many of the studies that we discuss employ the term GJT.

Closely related to AJTs are preference tasks, which ask participants to choose between two or more words, phrases, or sentences: While AJTs ask for a judgment of a given linguistic form, preference tasks ask participants to choose the best form for a given context. In contrast, the measures that we group under the umbrella term *interpretation tasks* do not ask participants to judge forms for grammatical acceptability at all. Instead, interpretation tasks ask participants to judge the truth of a sentence or to select a picture that matches the sentence; the participants' response informs the researchers about the status of particular linguistic forms in the learners' grammar.

This article provides an overview of recent studies that use AJTs, preference tasks, or interpretation tasks with adult and adolescent learners. The tasks discussed in this article can also be used with young children, with certain modifications; indeed, the types of interpretation tasks employed in SLA studies were originally created for use with young children. However, given our focus, we restrict our attention to task formats that are appropriate for use with cognitively mature, literate participants.

TRADITIONAL ACCEPTABILITY JUDGMENT TASKS

A traditional AJT presents sentences in isolation (with no preceding context) and asks participants to judge each sentence with regard to its grammatical acceptability. Some AJTs additionally ask participants to indicate their degree of certainty, also known as a confidence rating (see Rebuschat, 2013, for a discussion of this type of subjective measure). The sentences in a traditional AJT are designed to be either grammatically well-formed, or in violation of a rule of morphology, as in (1a), in which the plural marking on the object is missing, or syntax, as in (1b), which contains illicit movement of a subject *wh*-phrase. Such morphological and syntactic violations can be detected out of context; in contrast, semantic and pragmatic violations can only be evaluated in context, as discussed in a later section.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

(1) a. The farmer bought two pig at the market. [from Johnson & Newport, 1989]
b. Who did Sam say that likes Jane? [from White & Genesee, 1996]

SLA studies that use AJTs can be divided into two broad categories: (a) studies that focus on a limited set of linguistic structures, in order to examine how these structures are represented in interlanguage grammars; and (b) studies that test a variety of linguistic structures, in order to answer broader questions of linguistic proficiency, aptitude, or age effects. In addition to deciding on the number and type of structures to be tested, researchers must make at least four major decisions in constructing an AJT. These include modality of presentation (auditory vs. visual vs. bimodal); whether the task is timed or not; the type of response scale; and whether participants are asked to correct sentences that they judge as ungrammatical.

AJTS with a Specific Linguistic Focus

AJTs are very common in formal approaches to SLA, a consequence of the importance of acceptability judgments in formal syntax. Generative syntactic literature starting with Chomsky (1965, 1981) has treated native speakers' judgments of grammatical (un)acceptability as the main source of information about native speakers' linguistic competence. For an up-to-date overview of acceptability judgments in syntactic theory, see Sprouse (2013); see also Dąbrowska (2010) on judgments provided by linguist versus nonlinguist native speakers.

In a similar vein, SLA studies in the generative framework take second language (L2) learners' judgments of grammatical (un)acceptability to provide a window into learners' linguistic competence. AJTs used in generative approaches to SLA typically focus on a specific linguistic structure or set of related structures, in order to address particular theoretical questions. For example, a well-known study by White and Genesee (1996) used an AJT in which all experimental sentences were *wh*-questions. The goal of that study was to determine whether the grammars of adult L2 learners are constrained by Universal Grammar (UG), so all of the ungrammatical *wh*-questions in the AJT violated UG principles.

The AJT in White and Genesee (1996) used a binary response scale: each sentence had to be either accepted as grammatical or rejected as ungrammatical. Binary response scales for AJTs (labeled variously as yes/no, grammatical/ungrammatical, possible/impossible, for example) are quite common in SLA studies and are particularly appropriate when the contrast under consideration is clear-cut from the perspective of the native grammar. An example of such a clear-cut contrast is the placement of the object with respect to the verb in German, tested by Falk and Bardel (2011). This study used an AJT with a binary response scale to examine whether object placement in third language (L3) German was influenced by transfer from either of the learners' prior languages (English and French). Of those German sentences in the AJT that had incorrect object placement, half had word order that would be correct for English and

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

half had word order that would be correct for French. Therefore, incorrect acceptance of ungrammatical German sentences could ultimately provide information about whether learners were influenced by transfer from English, French, or both.

Falk and Bardel (2011) also asked participants to correct all sentences judged as ungrammatical; a response to an ungrammatical sentence was scored as accurate only if the learner rejected the sentence as well as provided an appropriate correction. This technique ensures that learners reject the sentences for the right reason. However, a potential downside of including an error correction component is that learners may overaccept sentences simply in order to avoid doing the extra work of error correction; alternatively, learners may recognize a sentence as ungrammatical, yet be uncertain about exactly where the error is, or about how to correct the error, and this, once again, may lead to overacceptance. And finally, focusing learners' attention on explicit error correction draws significantly on their explicit knowledge, whereas in many studies, AJTs are used with the goal of learning about the state of learners' underlying, implicit knowledge. All of these potential problems with error correction can be avoided if learners are first asked to provide their judgments (with no error correction) and afterwards are asked to provide corrections for those sentences that they rated as ungrammatical. This method was utilized by Gass and Alvarez-Torres (2005), who used an AJT on Spanish gender agreement and the copula *estar* as part of a pretest/posttest in an intervention study. In the AJT, participants read a series of Spanish sentences on a computer and rated each one as correct or incorrect. The AJT was timed, with each sentence appearing on the screen for only 18 seconds. At the completion of the AJT, participants received a printout of all the sentences they had marked as incorrect and were asked to provide error corrections.

Many SLA studies that use AJTs opt for Likert scales (typically with four or five, but sometimes as many as seven, response options) as an alternative to binary scales. For example, Bruhn de Garavito (2011) used a 5-point Likert scale (1 = ungrammatical, 5 = grammatical) in order to examine several different properties of subjects and objects in L2 Spanish. The results indicated that even for native speakers, not all types of ungrammatical sentences were considered to be equally unacceptable: Some types of ungrammatical sentences received mean ratings close to 1 on the Likert scale, but others had ratings closer to 3 (while still being rated lower than corresponding grammatical sentence types).

Whereas Bruhn de Garavito (2011) used a Likert scale consisting entirely of positive integers, other studies have used scales with both positive and negative integers. For example, Yuan and Dugarova (2012), investigating *wh*-topicalization in L2 Chinese, used a Likert scale from –3 (completely unacceptable) to +3 (completely acceptable). The use of negative and positive integers draws a clear dividing line between rejection (the response options of –3, –2, and –1) and acceptance (+1, +2, +3), in a way that a scale consisting exclusively of positive integers does not. However, the use of zero on such a scale is potentially problematic; in Yuan and Dugarova's study, as in many others, the response option of zero stood for *I don't know*. Such use of a zero midpoint conflates a scale of acceptability with a scale of certainty; some participants who choose

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

the zero option may be uncertain of their judgments, while others may be fairly certain that the sentence is neither very acceptable nor very unacceptable. A potential way around this problem would be to remove all zero responses from analysis; Yuan and Dugarova do not state whether they did this. The same problem (conflating acceptability with certainty) can potentially arise for an all-positive integer scale, especially one with an odd number of responses (e.g., on a scale of 1 to 5, the response of 3 may be interpreted by participants as *I don't know*). The problem can be alleviated by using a positive-integer scale with an even number of response options and/or by providing a "don't know" option that is not part of the scale. Montrul, Dias, and Santos (2011), in a study of clitic placement in L3 Brazilian Portuguese, implemented both of these solutions by using a scale from 1 to 4 plus a separate *don't know* response option, with all *don't know* responses subsequently discarded from analysis. The use of a Likert scale (as opposed to a binary scale) was motivated by an expectation of variability in native speaker responses to some of the constructions, which was then confirmed by the experimental findings.

The rating scales used in SLA studies, including those discussed earlier, are ordinal scales. However, as discussed by Cowart (1997), they are not necessarily interval scales (i.e., scales in which all intervals between adjacent response options are equal). Cowart points out that labeling each point on the scale increases the likelihood that intervals are not perceived as equal. For example, in Montrul et al. (2011), the numbers 1 through 4 on the scale corresponded, respectively, to *impossible*, *probably impossible*, *probably possible*, and *perfectly possible*. Participants are likely to perceive the interval between 2 (*probably impossible*) and 3 (*probably possible*) as greater than between 1 and 2 or between 3 and 4. While having an ordinal, noninterval scale is not in itself problematic, it does mean that parametric statistical tests (such as t-tests and analyses of variance [ANOVAs]) commonly used in SLA research are not as appropriate, since parametric tests assume that the data are interval data. Nonparametric tests such as the Chi-square test, the Mann-Whitney U test (the nonparametric equivalent of an independent samples t-test), or the Wilcoxon signed-rank test (the nonparametric equivalent of a paired samples t-test) can be used instead.

An alternative methodology to ordinal or interval rating scales is the magnitude estimation method (Bard, Robertson, & Sorace 1996), which originates in psychophysics. With this method, participants first give a rating to a modulus sentence (using any positive integer of their choice) and rate all subsequent sentences in proportion to the rating they gave to the modulus: For example, if a participant gives the rating of 100 to a modulus, a sentence that is twice as good as the modulus gets a rating of 200, while a sentence that is four times worse than the modulus gets a rating of 25. A recent study that uses magnitude estimation is Kraš (2011), which examined auxiliary change with unaccusatives in L2 Italian. Kraš converted the participants' ratings into analyzable data in several steps: First, each participant's numerical ratings were divided by the rating that this participant assigned to the modulus sentence (e.g., a score of 200 when the modulus is rated as 100 is identical to a score of 40 when the modulus is rated as 20). Next, the data were log-transformed, which ensured a normal distribution (satisfying an assumption of parametric tests). Finally,

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

differences between sentence pairs were computed for each participant, where the sentences in a pair differed mainly in the choice of auxiliary (*essere,* [be] versus *avere* [have]). The statistical analyses in Kraš were done on these differences in ratings, rather than on the ratings themselves, and used parametric tests (including a repeated-measures ANOVA and follow-up t-tests)

Magnitude estimation has several advantages over ordinal ratings scales, as discussed by Sorace (2010). First, the data obtained using magnitude estimation satisfy the assumptions of parametric tests, whereas data obtained from ordinal scales, which are often noninterval, do not. Second, ordinal scales often conflate acceptability with certainty, as discussed earlier. And third, ordinal scales require participants to provide absolute judgments of acceptability, which is difficult (in particular, as Sorace notes, participants may have difficulty interpreting the middle points on a multipoint scale); in contrast, magnitude estimation requires participants to judge the acceptability of sentences relative to other sentences, without having to provide absolute judgments. However, the advantages of the magnitude estimation method are challenged by Sprouse (2011), who showed (through experiments with native speakers) that the cognitive assumptions of magnitude estimation do not hold for acceptability ratings. Specifically, he argued that participants do not give ratio judgments, but may instead reduce the task to a standard multipoint rating scale. In light of this critique, the fact that most SLA studies that use AJTs rely on ordinal scales rather than on magnitude estimation does not appear problematic.

AJTS with a Broad Focus

Traditional AJTs have been used for decades in critical period research, which aims to clarify the effects of age on ultimate attainment in a second language as well as the nature of near-nativeness. (See DeKeyser, 2013, for a recent overview. It is worth noting that critical period researchers describe their instruments as GJTs rather than AJTs. However, we use the term AJT for consistency.) Critical period research typically relies on a correlational design in order to show the relationship between ultimate attainment—as measured by AJTs—and age of acquisition or other individual variables. One common thread linking the various types of critical period studies is the use of AJTs that are designed to test a wide range of structures. The number of structures tested varies anywhere from 3 to 12 or more, and structures are usually chosen based on their purported difficulty for L2 learners. As DeKeyser (2013) explains, "there needs to be a variety of structures so that generalization to (at least a well-defined subset of) the L2 grammar can be made" (p. 59–60). In other words, the AJTs in this line of research are broad by design in order to yield conclusions about participants' overall grammatical knowledge rather than their knowledge of one particular structure. This broad knowledge of the L2 is labeled in different ways across critical period research; for example, the AJT is used as a measure of proficiency, ultimate attainment in grammar, or grammatical intuition and processing.

A seminal study by Johnson and Newport (1989), one that has often been cited as evidence for critical period effects, tested participants' knowledge of English

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

syntax and morphology and correlated their performance to age of arrival in the United States. The main instrument was a traditional AJT that covered 12 areas of English morphosyntax (e.g., past tense, plural, third person singular, present progressive, determiners, pronominalization, particle movement). The test items were presented aurally and the response format was binary: grammatical or ungrammatical. Among the methodological limitations of the Johnson and Newport study was the length of the AJT (276 items), which may be seen as cognitively burdensome. Subsequent studies by DeKeyser (2000) and DeKeyser, Alfi-Shabtay, and Ravid (2010) used an abbreviated and modified version of the AJT in Johnson and Newport.

Another strand of critical period research aims to test the upper limits of SLA and therefore preselects individuals that pass as native speakers in everyday spoken communication. Accordingly, the AJTs used in this line of research need to be difficult in order to detect subtle differences between native and near-native speakers. For example, Abrahamsson and Hyltenstam (2008) constructed a "demanding" (p. 493) AJT in Swedish that included sentences that might pose difficulty even for native speakers. Using a similar design, Bylund, Abrahamsson, and Hyltenstam (2010) created an AJT in Spanish in order to describe the relationship between language aptitude and first language (L1) proficiency among speakers who may have suffered attrition (i.e., heritage speakers). The AJT tested three areas of Spanish morphosyntax: gender agreement, verbal agreement, and verbal clitics. As in previous studies, the AJT was presented aurally and the response format was binary (correct/incorrect). Bylund, Abrahamsson, and Hyltenstam (2012) used the same AJT as Bylund et al. (2010), but in combination with a cloze test that measured semantic and grammatical inferencing skills. Together, the two tasks (AJT and cloze) were intended as a measure of language proficiency, although the researchers analyzed the data from each task independently.

Two comments are in order regarding the methodology employed in the critical period studies described here. First, these studies report overall accuracy scores on the AJT in order to submit them to correlations with other variables. In other words, rather than analyzing or comparing scores on the different types of structures tested (as is the norm in AJTs with a specific focus), the relevant data for these studies is the overall score on the AJT (although see DeKeyser, 2000, for a critical period study that analyzed scores on different types of structures). Second, critical period researchers have generally favored the auditory format for AJTs. The issue of modality effects with AJTs was explored by Murphy (1997), who found that L2 learners were slower and less accurate in the aural modality and that this effect was most pronounced on ungrammatical items. Abrahamsson and Hyltenstam (2008) tested participants in both aural and written formats and found a strong positive correlation between the two modes ($r = 0.91$), which justified combining the scores from both AJTs. Note, however, that the L2 participants recruited by Abrahamsson and Hyltenstam were near-native. In this sense, the profile of the participants (e.g., their proficiency and the primary context in which they learned the language) is likely to interact with modality of presentation.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

In addition to critical period research, broadly constructed AJTs have also figured prominently in research on implicit and explicit knowledge. For example, the AJT in Ellis (2005) and in subsequent studies (Ellis et al., 2009) was designed to cover 17 morphosyntactic structures in English that are known to pose difficulty for L2 learners (e.g., modal verbs, adverb placement, dative alternation, verb complements). In a recent replication study of Ellis (2005), Bowles (2011) created a parallel AJT for Spanish that tested 17 challenging areas of Spanish morphosyntax (e.g., reflexive verbs, subjunctive, gender agreement, determiners, copular verbs). Since a wide range of structures was included, this means that there were relatively few tokens of each structure. The AJT was presented in written format on a computer screen and participants made a binary decision regarding grammaticality. Ellis (2005) and Bowles (2011) found that timed and untimed AJTs provide relatively independent measures of implicit and explicit knowledge. Both studies demonstrated that timed AJTs correlated with measures of implicit knowledge in contrast to untimed AJTs, which favored the utilization of explicit knowledge. If the interpretation of Ellis (2005) and Bowles (2011) is on the right track, then critical period studies (which generally use timed AJTs) have been measuring implicit knowledge. Furthermore, there may be a difference in how participants respond to grammatical versus ungrammatical items. In Ellis (2005), performance on the grammatical sentences on the untimed AJT correlated strongly with measures of implicit knowledge. In contrast, performance on the ungrammatical sentences correlated strongly with measures of explicit knowledge.

Finally, although the AJTs described in this section are broad in terms of the variety of structures tested, these are limited to morphosyntax. This follows from the fact that a binary scale is used to judge grammaticality, and thus test items cannot have more than one interpretation or be dependent on context. Given their relatively narrow focus on morphosyntax, researchers using these types of AJTs should caution readers that their findings cannot be generalized to all areas of the L2 (cf. DeKeyser et al., 2010).

PREFERENCE TASKS

An important variation of the traditional AJT is a task in which participants are asked to choose between two (or more) related sentences or forms in the target language. We will refer to this as a *preference task*, although the original studies referenced here have used a variety of labels (e.g., preference judgment task, sentence selection task, forced-choice task). Preference tasks can be used to test grammaticality (i.e., one form is grammatically correct while the other is not) as well as interpretation (i.e., one form expresses the target interpretation better than another). If interpretation is targeted, then the choice of form is generally preceded by some context. The nature of the response can vary as well; the contrast may be between words, phrases, or sentences. Finally, in any preference task, the researchers must decide whether the participants can choose only one option, both, or neither.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Sample Preference Tasks

To illustrate the use of preference tasks for the purpose of testing grammaticality, consider Zyzik and Azevedo (2009), a study that examined the ability of L2 learners of Spanish to distinguish between word classes. The preference task in Zyzik and Azevedo contrasted words belonging to the same family but differing in word class (e.g., *fuerte* "strong" vs. *fuerza* "strength"). In a sentential context, one of these forms is correct while the other is not. Participants were given the option of choosing one of the words, both, or neither. Zyzik and Azevedo designed the preference task based on contexts that were the source of errors in learners' written production, which allowed them to work with lexical items and constructions presumably familiar to the learners. For the purpose of studying word class distinctions, the preference task was a valuable methodology since it avoided the problem of metalanguage (e.g., noun, verb, adjective) and allowed competing forms to be presented side-by-side rather than in separate sentences.

An innovative use of preference tasks is to examine the acceptability of code-mixed discourse. Valenzuela et al. (2012) created a preference task in which participants (heritage speakers of Spanish) were asked to choose phrases and sentences that sounded most natural to them. The items on the preference task included a brief dialogue between bilingual speakers followed by a choice of determiner phrases that mixed a Spanish determiner and a noun in English (e.g., *la party* or *el party*). The question investigated by Valenzuela et al. was whether bilingual speakers would opt for the determiner that agrees with the inherent gender of the noun in Spanish or whether they would prefer the default masculine form. Thus, in the case of "party," the speaker might prefer the feminine determiner to reflect inherent gender *(la fiesta*—fem.*)* or alternatively choose the masculine determiner as a default. Since code-mixed discourse is difficult to characterize in terms of grammaticality, a preference task was a good option for investigating the nature of bilingual grammars in the Valenzuela et al. study.

Preference tasks are also a useful instrument for determining the interpretation of particular linguistic forms. A clear example is the case of Spanish copular verbs, *ser* and *estar* (to be), either of which can combine with the majority of adjectives, depending on the context and the speaker's intended meaning. Geeslin and Guijarro-Fuentes (2008) used a preference task to determine which variables are good predictors of copula choice among Spanish native speakers (studies on copula choice with L2 learners have used the same methodology: see Geeslin, 2003, 2005). An example from the preference task in Geeslin and Guijarro-Fuentes is given in (2).

(2) Paula and Raúl are planning to go out to dinner. Paula is yelling from the bedroom while she gets ready in order to make plans with Raúl. As she comes out of her room she asks:

Paula: Do you want to go in my car?

Raúl: *¡Ay! ¡Qué bonita estás! / ¡Qué bonita eres!* [How pretty you are!]

Note that the entire task was originally in Spanish and participants were instructed to choose one or both of the sentences to complete the dialogue.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Although either copula can be used with the adjective *bonita* ("pretty"), in the preceding context in (2), 70% of native speakers in the sample chose *estar*. The context contained several features that generally favor *estar* (e.g., individual frame of reference, immediate experience with the referent, susceptibility to change) and the task accurately reflected native speakers' preference for *estar*.

In Geeslin and Guijarro-Fuentes (2008), contexts were provided in the form of a written dialogue. However, preference tasks can also be paired with pictures to provide a context for interpretation. One recent example of this methodology is a study by Clahsen, Martzoukou, and Stavrakaki (2010), which investigated the acquisition of the perfective past tense in Greek. The materials for their preference task included sets of pictures; the first picture in each set depicted an ongoing activity (e.g., a child eating a cake) and the second picture showed the completed activity (e.g., an empty plate). The stimuli included existing and novel verbs with different kinds of morphological irregularities. Two versions of the preference task were administered: oral and written. In the oral task, participants heard two puppets (manipulated by the experimenter) produce two simple sentences to describe the second picture (note that the materials were originally used in Stavrakaki and Clahsen's 2009 study with young children). The two sentences were identical except for the past-tense form of the target verb, and participants were asked to choose which sentence was better, or alternatively, to provide a different past-tense form. In the written task, the two response options were printed below the picture and participants marked their preference. Since Clahsen et al. were interested in how L2 learners and native speakers generalize inflected forms to novel verbs, the pictures were necessary to establish the intended meaning of the novel verbs. Interestingly, Clahsen et al.'s comparison of the oral and written versions of the preference tasks revealed that the written task was more difficult for L2 learners, leading Clahsen et al. to suggest that the L2 learners were less adept than native speakers at using orthographic cues for making generalizations.

Acceptability Judgment Tasks and Preference Tasks Compared

The use of preference tasks versus traditional AJTs will depend on the specific goals of the research, but let us consider some advantages of preference tasks. First, in AJTs, it is impossible to know which aspect(s) of the sentence the participant is rejecting (unless the participants are instructed to make corrections). In contrast, preference tasks allow the researcher to make salient the contrast that is being targeted because the two sentences only differ in one aspect (e.g., past-tense form, use of copula, gender of determiner). In this way, there is a greater probability that learners focus on—and evaluate—the feature that the researcher is interested in. Another advantage of preference tasks is that they are well suited for studying variable phenomena. We saw the example of copula choice (Geeslin & Guijarro-Fuentes, 2008) in which one copula is generally preferred in certain contexts, but there is expected variation among native speakers. Finally, preference tasks allow the researcher to manipulate the language (L1 versus L2) of the context. When the population consists of more advanced learners or native speakers, contexts in the target language

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

are appropriate. Yet presenting the contexts in the participants' L1 is a good option when working with lower-proficiency learners to ensure that the context is unambiguously understood (see Dekydtspotter, Sprouse, & Anderson, 1997 for a detailed discussion of this issue). Note that context-based AJTs also allow the researcher to choose either L1 or L2 for the context (see the next section).

At the same time, preference tasks have some disadvantages relative to AJTs. First, by focusing on the contrast between the two forms, preference tasks make the contrast very explicit and increase the probability that learners will draw on their metalinguistic knowledge rather than on their intuitions about grammar. This is, of course, only a problem if the researchers' goal is to target underlying intuitions and implicit knowledge rather than explicit, metalinguistic knowledge. Another potential limitation of a preference task is that it only tells us that one form is preferred to another, but does not inform us of whether the nonpreferred form is unacceptable, or acceptable but only nonpreferred. Ultimately, the choice of task should depend on the researchers' goals, as well as on a consideration of learners' abilities.

CONTEXT-BASED ACCEPTABILITY JUDGMENT TASKS

As discussed earlier, traditional AJTs present sentences in isolation, with no preceding context. This format is appropriate for testing morphological phenomena (such as agreement marking on a verb or noun) and local syntactic phenomena (such as word order in a *wh*-question), where ungrammaticality is detected by native speakers without any context. However, many other linguistic phenomena cannot be judged in isolation, requiring prior context. The nature of the context differs across studies, ranging from a single sentence to a text passage and/or a picture, and different studies use different names for judgment tasks that include contexts (e.g., sentence continuation judgment task, sentence-pair interpretation task). We group all such tasks under the broad category of context-based AJTs. Depending on the linguistic phenomenon and the nature of the context, the instructions to participants differ as well: while some context-based AJTs ask participants to judge acceptability, others ask for judgments of naturalness, appropriateness, or felicity, as discussed in the following section.

Context-based AJTs That Test Grammatical Acceptability and Unacceptability

An example of a syntactic phenomenon that cannot be judged without prior context is verb phrase (VP) ellipsis, tested by Duffield and Matsuo (2009) as well as Duffield, Matsuo, and Roberts (2009); both studies built on the design of Tanenhaus & Carlson (1990). A sentence such as *Tom told us that Sally did* is grammatical, but is not interpretable without preceding context in that *did* here stands in for an elided VP and context is needed to recover the VP. In order to test L2 English learners' sensitivity to the properties of VP-ellipsis, Duffield and Matsuo (2009) presented participants with sentence pairs such as those in (3), where the first sentence provides the antecedent for the elided VP in the

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

second sentence. Participants had to judge whether the second sentence was acceptable as a continuation of the first.

(3) a. Someone took the wood out to the shed last night. Tom told us that Sally did.
b. The wood was taken out to the shed last night. Tom told us that Sally did.

One of the questions tested in this study was whether L2 learners, like native speakers, would find VP-ellipsis less acceptable when the VP antecedent was in the passive voice (as in [3b]) than when it was in the active voice (as in [3a]; the study tested several other related constructions as well, including VP-anaphora, as in *Tom told us that Sally did it*). The task was timed, providing information about response time as well as acceptability: In general, continuations with lower acceptability (such as [3b]) yielded longer response times than their counterparts with higher acceptability (such as [3a]). In a follow-up study, Duffield et al. (2009) tested additional syntactic properties of VP-ellipsis and also introduced a methodological change by using both timed and untimed versions of the task: In the offline task, participants judged the acceptability of the second sentence in the context of the first on a scale from 1 to 5, whereas in the online task, they provided a binary judgment of acceptability, under time pressure. Despite these differences, the timed and untimed tasks provided convergent results.

Another study that tested both grammatical and ungrammatical sentences in context is Year and Gordon (2009). Working within a usage-based approach, the researchers conducted a training experiment with the English ditransitive construction. Participants were middle school students who spoke Korean as their L1. During the training session, video animations were used to teach participants prototypical three-argument verbs such as *give* that can appear in both the ditransitive construction (e.g., *Peter gave Karen a book*) and the *to*-prepositional construction (*Peter gave a book to Karen*). Furthermore, participants were taught verbs that are generally unacceptable in the ditransitive construction (e.g., *donate*) as well as novel verbs that mirrored the properties of either *give* or *donate*. To assess learning, a context-based AJT was used as one of the tasks. The context was provided with video clips that depicted simple events with an agent, a recipient, and an object being transferred. The experimental sentences were presented in written form (at the bottom of the screen) and also aurally. Participants were asked to judge the acceptability of the sentences on a 7-point Likert scale with a midpoint value (4) for the option *I don't know*. The experimental sentences included different construction types (ditransitive and *to*-prepositional) using both real and novel verbs. Although sentences such as *Peter gave Karen a book* could, in theory, be evaluated in isolation, several aspects of the study required contextual support. First, the video animations provided an unambiguous meaning for the test sentences; this was especially important given the inclusion of novel verbs that would have otherwise yielded incomprehensible sentences (e.g., *Peter greemed Karen some candy*). Additionally, the young age of the participants generally favored the use of pictures and/or video context.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Another study that used context in combination with ungrammatical test sentences is Zyzik (2014), which focused on the syntax-lexical semantics interface. Specifically, this study examined Spanish heritage speakers' acceptance of intransitive verbs in transitive sentences (i.e., argument-structure overgeneralizations). Strictly speaking, an intransitive verb such as *crecer* ("grow") is ungrammatical in a transitive sentence, but Zyzik hypothesized that heritage speakers might accept such overgeneralizations due to transfer from English (note that "grow" in English can be used transitively). The main instrument in Zyzik (2014) was an untimed AJT in which each item consisted of a short context (in Spanish) followed by a test sentence that participants rated on a 4-point acceptability scale. If participants rated an item as a 1 or a 2, they were asked to fix the sentence to make it acceptable. The contexts were designed to cue participants to a causative meaning, as illustrated in (4).

(4) Context: *La madre de Pablo lo llevó al dentista. Mientras estaba en la sala de espera, escuchaba los llantos de su hijo.* "Pablo's mother took him to the dentist. While she was in the waiting room, she could hear the cries of her son."

Test sentence:

El	*dentista*	*lloró*	*a*	*Pablo.*
the	*dentist*	*cried*	ACC	*Pablo*

"The dentist cried Pablo."

Many of the test sentences could have been ambiguous without the intended causative meaning provided by the context since some intransitive verbs (e.g., *correr* ["run"]) have a transitive use in some varieties of Spanish. In earlier studies with L2 learners (cf. Montrul, 1999), pictures were used to create such contexts and in research with young children (cf. Ambridge, Pine, Rowland, & Young, 2008), computer animations were used for the same purpose. The insight is that sentences may not always be interpreted as intended by the researchers, and thus, even if a sentence is ungrammatical (as in the case of argument-structure overgeneralizations), it is a good idea to provide contexts in order to ensure that all participants are evaluating the same meaning.

Context-Based AJTs that Test Interpretation

In the studies described in the previous section, the test sentences are ultimately either grammatical or ungrammatical, but the grammaticality or ungrammaticality of the sentence needs to be evaluated in context. In many other studies, all test sentences are grammatical, but not all sentences are equally acceptable in all test contexts. An example is Ivanov (2012), a study that used a context-based AJT to examine whether learners had acquired the pragmatic function of clitic doubling in Bulgarian as a marker of topicality (see also Slabakova, Kempchinsky, & Rothman, 2012, on clitics in L2 Spanish). Each test item in Ivanov's study consisted of a text passage in English (the learners' L1), providing the context, and four Bulgarian sentences that differed along two dimensions (whether the object noun phrase [NP] was fronted and whether it was doubled by a clitic), but were identical in all other respects. All four sentences were grammatical out

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

of context, but were either pragmatically felicitous or pragmatically infelicitous in the context provided, depending on whether the object NP was the topic or the focus. Participants were instructed to rate each sentence on a scale from 1 to 5 for its acceptability in the context. Presenting all four target sentence types side-by-side after each story cut down on the length of the test but had the potential downside of drawing participants' attention to the target contrast, causing participants to make explicit comparisons across sentence types. Presenting the context in the learners' L1 ensured that the context would be fully understood.

Similarly to Ivanov (2012), Zhao (2012) tested a phenomenon at the syntax/pragmatics interface (the interpretation of null and overt arguments in L2 Chinese) and presented test sentences side-by-side within each test item. The contexts in Zhao (2012) were expressed by means of pictures rather than text; use of pictures, like use of passages in the learners' L1, ensures full understanding of the context. Each picture showed two people, with one of them making a statement (in a speech bubble) either about him- or herself or about the other person in the picture (e.g., *Grandpa Lu likes watching football matches*). The picture was accompanied by four sentences: two test sentences and two distracters. The test sentences differed minimally in containing either an embedded overt pronoun or an embedded null pronoun. Overt and null pronouns in Chinese differ in their ability to refer to the matrix subject versus to an outside referent. For example, if a learner rates *Maomao says that he likes football matches* as acceptable in the context of a picture where Maomao is talking about his grandpa liking football, this indicates that the learner (correctly) allows *he* to refer to an outside referent that is not the matrix subject. Participants rated sentences on a scale from –2 to +2, where zero was replaced by an *I don't know* option. Additionally, a response option of *incorrect sentence* was included: If participants found null pronouns in Chinese to be ungrammatical (as are null pronouns in English, the learners' L1), they would be expected to choose this option. This control ensured that participants were only judging (for interpretation) sentences that they considered grammatical. Both *I don't know* and *incorrect sentence* responses were discarded from further analysis. While replacing zero with "I don't know" avoids the problem of conflating acceptability with certainty (discussed earlier), this replacement means that the interval between two adjacent points of the scale (–1 and +1) is twice as large as the intervals between other pairs of adjacent points (–2 and –1; +1 and +2). This could easily be adjusted for statistical analysis, but it is not clear whether that was in fact done.

INTERPRETATION TASKS

Interpretation tasks can be divided into three broad categories. The first category consists of tasks in which participants are provided with a context (in the form of a text passage, a picture, a video, or a story acted out with props) that establishes a particular interpretation; participants are then presented with a target sentence and have to make a judgment of that sentence given the preceding context. Context-based AJTs fall into this category, as do truth-value

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

judgment tasks (TVJTs). The second category consists of tasks in which the procedure is reversed: Participants are first provided with the target sentence and are asked to choose a particular interpretation for that sentence from among two or more options, which are presented as either text or pictures. The third category consists of tasks in which participants are asked to carry out a command or request; in its classic form, this is known as the act-out task.

Most interpretation tasks used in L2 acquisition research were originally designed for use in L1 acquisition research. The interpretation tasks used in L1 acquisition research are nearly always oral rather than written in format and often include toys and puppets to provide the context and/or interact with the children. For a detailed discussion of interpretation tasks in L1 acquisition, see the edited volume by McDaniel, McKee, and Smith Cairns (1996); see also Crain and Thornton (1998) on TVJTs. For a recent overview of interpretation tasks used in L1 acquisition research, see Schmitt and Miller (2010), and for a similar overview for L2 acquisition, see Ionin (2012).

Sample TVJTs

When interpretation tasks originally designed for use with children are adapted for use with literate, adult L2 learners, the format often changes from oral to written, and a story acted out with toys is changed into a story presented using text, possibly accompanied by pictures. For example, Grüter, Lieberman, and Gualmini (2010) made such changes when adapting the TVJT used with young children by Goro and Akiba (2004). Both studies examined the scope of disjunction in negative sentences such as (5). In English, (5) conveys that the pig ate neither the carrot, nor the pepper: the disjunction *or* takes narrow scope with respect to negation, so that (5) can be paraphrased as "It is not the case that the pig ate the carrot or that the pig ate the pepper." In contrast, in Japanese, disjunction takes scope over negation, so that (5) can be paraphrased as "The pig didn't eat the carrot, or the pig didn't eat the pepper." Thus, in a context where the pig ate the carrot but not the pepper (or vice-versa), (5) is false in English, but its equivalent in Japanese is true.

(5) The pig didn't eat the carrot or the pepper.

Prior studies on L1 acquisition used the TVJT format to investigate how young children interpret sentences such as (5) in English (Gualmini & Crain, 2005) and Japanese (Goro & Akiba, 2004). Grüter et al. (2010) extended the investigation to L1 English L2 learners of Japanese as well as L1 Japanese L2 learners of English. In the TVJT used by Grüter et al. (2010), participants read a short story, accompanied by pictures, about an eating contest among several animals that involved eating a cake, a carrot, and a pepper. Then participants read statements made by other characters (e.g., Hello Kitty) about the outcome of the eating contest and marked the statements as right or wrong. For example, in one condition, the pig received a prize that indicated that it ate one of the vegetables; if Hello Kitty subsequently uttered (5) to state what happened, she would be wrong in English (since in English, [5] means that the pig ate neither the carrot nor the pepper), but correct in Japanese (since in Japanese, [5] means

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

that the pig either did not eat the carrot or did not eat the pepper—but still ate the other vegetable). The learners' evaluation of Hello Kitty's utterance as right or wrong thus informed the researchers about how they interpreted the scope of disjunction in their L2.

The written format, in which participants marked *right* or *wrong* on the page next to the picture of Hello Kitty (rather telling their response to a Hello Kitty puppet), made the study appropriate for adult learners; in other respects, Grüter et al.'s (2010) study was very similar in design to TVJTs used with young children. In contrast, many other SLA studies use modified TVJTs (which are similar in format to context-based AJTs) in which participants rate the sentence on scale for how well it matches the preceding context. Studies that use modified TVJTs include Marsden (2009), a study of quantifier scope in L2-acquisition, as well as Gabriele (2009), a study of aspectual interpretation.

Gabriele (2009) examined the interpretation of progressive aspect for L1 English–L2 Japanese and L1 Japanese–L2 English learners. Several different aspectual forms were tested, including the progressive aspect with achievement verbs such as *arrive*, where English and Japanese behave differently. In English (6) has a progressive interpretation: The plane is in the process of arriving at the airport. In Japanese, the imperfective marker *te-iru*, combined with an achievement verb, yields a resultative interpretation, so that the Japanese literal equivalent of (6) means that the plane has arrived and is at the airport.

(6) The plane is arriving at the airport.

In order to examine transfer in this domain, Gabriele (2009) used a story compatibility task that was essentially a modified TVJT. The contexts were presented by means of pictures plus audiorecorded narration; the test sentence appeared on the screen after the story had been presented. There was a complete story version (for [6], the story made it clear that the plane was already at the airport) and an incomplete story version (which indicated that the plane was still in the air). The English sentence in (6) is false/unacceptable after the complete story, but true/acceptable after the incomplete story; for the Japanese equivalent of (6), the judgments are reversed. Participants rated the sentences on a scale from 1 (*I definitely cannot say this sentence in the context of the story*) to 5 (*I definitely can say this sentence in the context of the story*). The midpoint, 3, was defined as "I might be able to say this sentence in the context of the story." This suggests that 3 was not a true midpoint, but was skewed toward an acceptance response. The participants also had the option of leaving an item blank if they were not sure of their response.

As just seen, the TVJT format is particularly amenable for the testing of scope and of aspectual interpretation. Another domain in which TVJTs are commonly used is anaphor binding. It is well known that in English, reflexives such as *herself/himself* must be locally bound: for example, in (7), *herself* must be co-indexed with *Sue* (the local antecedent) and not *Mary* (the long-distance antecedent). In contrast, reflexives in many other languages, including Chinese and Korean, do allow (and for some reflexive pronouns, prefer) long-distance antecedents.

(7) Mary_i thinks that Sue_j painted $\text{herself}_{j/*i}$.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Much research in both L1 acquisition (e.g., Chien & Wexler, 1990, and much subsequent work) as well as L2-acquisition (e.g., Finer & Broselow, 1986) has investigated how learners interpret reflexives in the target language, and, in the case of L2 learners, whether they are influenced by the interpretation of reflexives in their L1. A variety of methodologies have been used to study this phenomenon, including act-out tasks (e.g., in Chien & Wexler, 1990, children had to respond to a command such as *Kitty says that Sarah should point to herself*, where Sarah is the child's name); forced-choice preference tasks (e.g., in Thomas, 1991, L2 learners had to choose whom *herself* refers to from the options provided); and most commonly, TVJTs.

A recent SLA study on this topic is Kim, Montrul, and Yoon (2009), which used a picture-based TVJT (based on White, Hirakawa, & Kawasaki, 1996) to examine how Korean heritage speakers judge different reflexive pronouns in Korean. For example, the Korean form of a sentence like *Cheli said that Minswu drew himself* would be presented either with a picture in which Cheli talks about Minswu drawing Cheli (long-distance antecedent) or with a picture in which Cheli talks about Minswu drawing Minswu (local antecedent). Speech bubbles were used to show speech, and the characters in the pictures wore labels indicating their names. Participants judged whether the sentence was true or false in the context of the picture. Use of a picture instead of a story to present the context ensures that participants understand the context; another approach is to present the story context in the learners' L1, as was done, for example, in Jiang (2009), another study that used a TVJT to study anaphor binding in L2 acquisition.

Other Types of Interpretation Tasks

The TVJT (typically modified in some way) appears to be the most common type of task used in L2 acquisition studies that examine semantic or pragmatic phenomena. Another task type that is relatively common is a forced-choice interpretation task. Unlike the forced-choice preference tasks described earlier, which ask participants to choose the best sentence or form for a given context, forced-choice interpretation tasks provide the form and ask participants to choose the best interpretation for it. Such tasks are commonly used to study anaphora resolution, that is, how speakers resolve the ambiguity in sentences such as (8).

(8) John saw Charles when he was walking on the beach (Keating, VanPatten, & Jegerski, 2011, p. 197)

The crux of the problem is which antecedent (John or Charles?) is linked to the subject pronoun (*he*) of the subordinate clause in (8). Anaphora resolution is complicated in null-subject languages since the subject of the subordinate clause can be either overt of null. One way of investigating participants' interpretation is with a forced-choice interpretation task in which the stimulus sentence (such as [8]) is followed by a comprehension question (e.g., *Who was walking on the beach?*) with two options that correspond to the possible antecedents. In Keating et al. (2011) participants saw the stimulus sentence (in Spanish) on a computer

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

screen, followed by the comprehension question and response options on a separate screen. The task was untimed and participants were asked to make a binary decision. Jegerski, VanPatten, and Keating (2011) used the same task format to study anaphora resolution in both L2 English and L2 Spanish.

A forced-choice interpretation task with pictures, also known as a picture-matching task (PMT), was used by Papadopoulou et al. (2011) to study case marking and word order in the L2 Turkish of L1 Greek speakers. Participants heard a sentence such as *The man is embracing the woman* in either subject-object-verb (SOV) or object-subject-verb (OSV) order; accusative case marking on *the woman* was a cue that this was the object, regardless of word order. Participants had to match the sentence to one of four pictures, two of which depicted the same action but with reversed thematic roles (a picture of a man embracing a woman vs. a picture of a woman embracing a man; the other two pictures were distracters); if participants relied on word order rather than case marking to assign thematic roles, they would choose the wrong picture in response to an OSV sentence. In addition to the PMT, Papadopoulou et al. used a traditional AJT to investigate whether learners were aware that accusative-marked objects can appear before the subject, while non-case-marked objects cannot.

It is important to mention that PMTs similar to the ones in Papadopoulou et al. (2011) are a useful tool in instructed SLA research. For example, PMTs have figured prominently in research on Processing Instruction (PI), dating back to VanPatten and Cadierno (1993). Several PI studies have targeted Spanish clitic pronouns, which can appear in sentence-initial position, yielding object-verb-subject (OVS) word order. The PMT presents participants with an aural stimulus such as (9), and participants are asked to choose one of two pictures that represents the event.

(9) Lo ve la profesora.
Him sees the professor (fem.).
"The professor (f.) sees him."

Given the example in (9), the two pictures would depict (a) a man looking at a female professor (incorrect) and (b) a female professor looking at a man (correct). If participants process the sentence based on English word order cues, they will misinterpret OVS sentences with clitic pronouns. Accordingly, PI aims to correct learners' inefficient processing strategies by manipulating the input in such a way that learners are pushed to pay attention to formal cues (e.g., case marking, verbal inflections) in order to get the meaning. Since PI is a pedagogical intervention, studies of this sort often use a pretest/posttest design to measure learners' improvement over time; PMTs are the assessment measure used at the time of pre- and posttesting. Some recent research has begun to use a trails-to-criterion method; that is, the number of items presented before the learner begins to process the input correctly (cf. Fernández, 2008; Henry, Culman, & VanPatten, 2009). For example, VanPatten and Borst (2012) examined the problem of Spanish clitic pronouns illustrated in (9) and determined that processing three OVS sentences followed by one SOV distracter

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

would constitute criterion. The results of the study, thus, do not compare which experimental group registered greater improvement, but rather which group begins to process the sentences correctly earlier. One advantage of using the trials-to-criterion method in intervention research is that different versions of the PMT do not have to be created for pre- and posttests.

Unlike TVJTs and forced-choice interpretation tasks, act-out tasks are not commonly used in research with L2 learners; a likely reason for this is that the act-out format resembles a game and appears to be more appropriate for studies with young children than with adults. Ionin, Baek, Kim, Ko, and Wexler (2012) used a modified act-out task with adult L2 learners, in which participants carried out commands to draw various geometric shapes, as in (10).

(10) Here are six cars and six books.
1. Please draw arrows above two books.
2. Now, please draw circles around the books.

This study examined whether speakers of Korean, an article-less L1, assign the same interpretation for "the" as do native English speakers. In line 2 in (10), *the books* can potentially denote either all six books mentioned at the beginning, or just the two books above which the participant drew arrows in response to line 1. For native English speakers, both readings are available to *the books*; in contrast, the demonstrative phrase *those books*, used in the same context, must necessarily denote the two salient books, not the entire set of six books. The study examined the interpretation that L1 Korean–L2 English learners assigned to *the books* and whether learners' performance differed from that of native speakers.

Context-Based AJTs Versus TVJTs: The Case of Plural NP Interpretation

A single phenomenon can potentially be investigated using a variety of different tasks. For example, Hwang and Lardiere (2013) investigated different properties of plural marking in Korean using a traditional AJT, a preference task, and a TVJT. As discussed earlier, anaphor binding has been investigated using TVJTs, forced-choice tasks, and act-out tasks. Another example of a linguistic phenomenon that has been tested by means of different interpretation tasks is the interpretation of plural NPs as generic versus specific.

The basic phenomenon under investigation has to do with whether a plural NP requires an article to express generic meaning. In English and German, the generic interpretation is available to bare plural NPs (*Lions are dangerous* is a statement about lions in general); plural NPs that occur with a definite article have only a specific, nongeneric interpretation (*The lions are dangerous* is a statement about a specific group of lions, not about lions in general). Conversely, in Spanish and Italian, a definite article is required for generic as well as specific interpretations.

Several recent studies have investigated the role of transfer in this domain, with bilinguals who speak English or German as one of their languages and Spanish or Italian as the other. Serratrice, Sorace, Filiaci, and Baldo (2009) tested

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

bilingual English/Italian children on NP interpretation in both English and Italian, using a context-based AJT. The test sentences, which were presented aurally, came in the four conditions illustrated in (11), where the adverbial *here* sets up a specific interpretation, while *in general* sets up a generic interpretation. The sentences were accompanied by pictures of prototypical objects (red strawberries or dangerous sharks). Note that despite the presence of pictures, this AJT tested (un)grammaticality rather than appropriateness: The participants had to say whether the sentence sounded acceptable in English/Italian.

(11) a. Here, strawberries are red.
b. Here, the strawberries are red.
c. In general, sharks are dangerous.
d. In general, the sharks are dangerous.

Arguably, the English and Italian versions of the task were not fully comparable. In Italian, the sentences with bare plurals, (11a) and (11c), are ungrammatical and the sentences with definite plurals, (11b) and (11d), are grammatical; in order to provide a response, participants had to simply note whether the plural NP was preceded by a definite article. In the English version, however, participants needed to pay attention to the adverb as well as the article: For the specific interpretation, established by *here*, the definite plural is required ([11b] is acceptable, [11a] is not); for the generic interpretation, established by *in general*, the bare plural is required ([11c] is acceptable, [11d] is not). Serratrice et al. found that in English, the child and, to a lesser extent, even the control adult participants did not always take the adverbial into consideration when giving their responses.

Other studies on the generic versus specific interpretation of plural NPs have used interpretation tasks rather than (or in addition to) judgment tasks. Ionin, Montrul, and Crivos (2013) adapted the TVJT methodology used by Pérez-Leroux, Munn, Schmitt, and Delrish (2004) with L1-acquiring children to a study with adult L1 Spanish–L2 English and L1 English–L2 Spanish learners (the former population is also discussed in detail in Ionin & Montrul, 2010). Participants read short stories about two unusual representatives of various animal species: tigers who are vegetarian, birds who live in caves, and so on. Each story was accompanied by a picture of the unusual animals. Participants then had to judge the target sentence as true or false in the context of the story and picture. In English, the sentence *Tigers like meat* is true in the context of the story, since tigers in general like meat. In contrast, *The tigers like meat* is false, since the particular tigers in the story are vegetarian. In contrast, the Spanish equivalent of *The tigers like meat* is true if *the tigers* is interpreted generically and false if it is interpreted specifically. Thus, participants' response of true versus false informed the researchers of whether learners were assigning the target interpretation to the definite plural, or transferring the interpretation available in their L1. However, the English and Spanish versions of the study were not directly comparable. In English, *the tigers* has only a specific interpretation, and a response indicating a generic interpretation is simply incorrect. In contrast,

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

in Spanish, *the tigers* has both generic and specific interpretations, and indeed both interpretations were allowed by native Spanish speaker controls. The differences between native Spanish speakers and L2 Spanish learners were a matter of preferences rather than absolute judgments: Native speakers preferred the generic interpretation while lower-proficiency learners preferred the specific interpretation. In order to get at judgments as well as preferences, Ionin et al. conducted an additional study using a context-based AJT in which participants read short (single-paragraph) stories that established either a generic or a specific interpretation and rated the target sentences for (un)acceptability in the context, using a scale from 1 to 4. Multiple test sentences (variants with both bare and definite plurals, as well as with singular NPs) were presented side by side, which made the task more explicit; overall, learners were more target-like in the AJT than in the TVJT, possibly as a result of the more explicit nature of the former.

Kupisch (2012) also employed both TVJT and AJT methodologies in a study with adult L1 German–L2 Italian learners as well as adult simultaneous German/Italian bilinguals (the participants were all tested on NP interpretation in Italian). For the TVJT, Kupisch used only pictures (no stories) to present the context; each picture was followed by three different sentences (the target sentence and two controls), presented aurally. For example, a picture of tailless kangaroos would be accompanied by the target sentence *The kangaroos have tails*, in Italian. If participants interpreted *the kangaroos* as specific (possible in both German and Italian), they would respond *false* to the sentence in the context of the picture; if they interpreted it as generic (possible in Italian but not in German), they would respond *true*. Kupisch noted the issue (discussed earlier) of the TVJT getting a preferences rather than absolute judgments, since definite plurals in Italian have both generic and specific readings. Kupisch also used a context-based AJT to examine whether learners considered definite plurals with generic versus specific interpretations to be acceptable in Italian. Each item consisted of a pair of sentences, with the first sentence establishing either a generic or a specific context and the second sentence containing a definite or a bare NP. The sentences were presented both visually and aurally, and participants were under time pressure to respond and to correct the second sentence if they judged it as ungrammatical.

To sum up, the three preceding studies used two types of tasks—context-based AJTs and TVJTs—with variation both in the modality of the presentation and in how the context was presented. For the AJT, the context was established by a single adverbial phrase (Serratrice et al., 2009), a single prior sentence (Kupisch, 2012), or a paragraph (Ionin et al., 2013). In the TVJT, the context was established by a story plus a picture (Ionin et al.) or a picture alone (Kupisch). The tasks further differed in whether multiple sentences were presented side by side after each context, versus one at a time. All of these differences could potentially influence learners' responses: For example, Kupisch (2012) noted that presenting the TVJT context using only pictures of unusual animals may bias the participants toward a specific interpretation; in contrast, some features of the TVJT design in Ionin et al. (2013) may have biased the participants toward a generic interpretation. Importantly, whatever the bias of a specific task, any

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

differences obtained among different populations tested with that particular task are still valid. However, the methodological differences between studies make direct comparisons across studies difficult.

Summary: Interpretation Tasks

To summarize, SLA researchers have a variety of tasks at their disposal for investigating how learners interpret phrases and sentences in their L2. All tasks have specific advantages and disadvantages. A TVJT has an advantage over a context-based AJT in that it focuses on meaning rather than form and is more implicit; however, at least for some phenomena, TVJTs may inform researchers only about preferences for a particular interpretation and not about absolute judgments of which interpretation(s) are available. A PMT or another type of forced-choice interpretation task similarly informs researchers about which interpretation is the most preferred. A context-based AJT gets most directly at learners' judgments of a given interpretation; however, this task type is also the most explicit, in that it asks learners to think about the grammatical acceptability of a sentence and not just about its meaning. Auditory presentation and time pressure are some ways in which tasks can be made more implicit.

For any given task, such as a TVJT, there are many differences across studies in terms of how the context and the response options are presented. Possible options include responses of *true* versus *false*, *right* versus *wrong*, *matching* versus *not matching* (to the story), or *can be said* versus *can't be said* (in the context of the story). The responses can be limited to just two options, or to a scale with multiple options, exactly as in the case of an AJT. The context can be presented as a passage of text, a picture, an acted-out story, or some combination of these. An interesting empirical question is whether, or how, the nature of the context, the type of response, and the number of points on the response scale, influence participants' performance.

CONCLUSION

This article has presented a broad overview of recent research in SLA that measures learners' linguistic knowledge without eliciting production of the target language. We have surveyed four main types of tasks: traditional AJTs, preference tasks, context-based AJTs, and interpretation tasks. As evident from the wide range of studies included here, these tasks form a heterogeneous category that requires many layers of decision-making on the part of the researcher. One of the primary considerations is the nature of the linguistic phenomenon that one wishes to target. Specifically, the researcher must determine whether the linguistic contrast can be evaluated out of context. If so, then a traditional AJT with a binary response scale may be appropriate. Traditional AJTs are typically limited to the study of morphosyntax, but can be used to address very different research questions—those that focus on the status of a particular construction in the learners' interlanguage and those that address broader issues of proficiency and ultimate attainment.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

If the linguistic phenomenon of interest is dependent on context or includes novel forms (cf. Clahsen et al., 2010; Year & Gordon, 2009), then a context-based AJT or interpretation task should be considered. Semantic and pragmatic phenomena require context: The same grammatical sentence may be appropriate and felicitous in one context, but inappropriate or infelicitous in another. The inclusion of context makes for another level of decision-making, including the type of context (sentential, paragraph, or picture) that will clearly depict the intended meaning. Moreover, the language of the context (L1 vs. L2) is an important factor that must be considered in light of the proficiency level of the learners.

Finally, there are methodological choices to be made independent of the question of contextual support. These include the modality of presentation, time pressure, and type of response scale. We have outlined some of these considerations in detail, as well as some of the pitfalls that can negatively affect the validity and/or reliability of the task (e.g., conflating scales of acceptability and certainty). Although many variables can be controlled or minimized with careful task design, some learner variables are more invisible and therefore difficult to gauge. One such variable is the question of whether a learner is utilizing implicit knowledge in rejecting or accepting a target sentence, or conversely, if explicit (and possibly metalinguistic) knowledge is being relied on in the process. Certain tasks, such as preference tasks, make linguistic contrasts more explicit and may predispose learners to focus on form rather than meaning.

Taken together, AJTs and interpretation tasks provide a valuable set of methodological options for SLA researchers working with various theoretical frameworks. Although AJTs are typically associated with formal or generative approaches, our survey of the research has shown that AJTs are not necessarily bound to a single theoretical perspective. Indeed, some usage-based studies (e.g., Year & Gordon, 2009) have incorporated AJTs in their design. Likewise, instructional intervention studies (e.g., VanPatten & Borst, 2012) that focus on processing rather than syntactic representation have benefitted from interpretation tasks. Finally, we have shown that many of these tasks can be used in combination with one another (e.g., context-based AJTs and TVJTs) to yield a more complete and insightful picture of learners' knowledge. And finally, all types of judgment and interpretation tasks can be used in conjunction with production data, allowing researchers to examine whether the same or different factors are at work in learners' production and comprehension.

ANNOTATED BIBLIOGRAPHY

Cowart, W. (1997). *Experimental syntax*. Thousand Oaks, CA: Sage.

This textbook covers the construction, administration, and analysis of acceptability judgment tasks that test syntactic phenomena. Although the book focuses exclusively on experimental syntactic research with native speakers, the content is fully applicable to research with second language learners. The topics covered include a detailed discussion of experimental design and an introduction to statistical analysis.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Dąbrowska, E. (2010). Naïve versus expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*, 1–23.

This article discusses many of the methodological issues associated with AJTs as necessary background to an empirical study in which linguists (of both generative and functional orientations) were compared to nonlinguists in their evaluation of sentences containing long-distance dependencies. The results indicate that linguists' judgments diverge from those of the general population. Dąbrowska suggests that this is due to linguists' frequent exposure to specific kinds of ungrammatical sentences.

Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, *27*, 141–172.

In this article Ellis describes a battery of tests, including two versions of an acceptability judgment task, that were designed to provide relatively independent measures of implicit and explicit knowledge. The results of psychometric testing indicate that timed and untimed AJTs measure different types of knowledge; time pressure seems to encourage participants to rely on their implicit knowledge while an untimed AJT is more likely to tap explicit knowledge.

Ionin, T. (2012). Formal theory-based methodologies. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 30–52). Oxford, UK: Wiley-Blackwell.

This book chapter covers the use of judgment and interpretation tasks in formal, generative research with adult second language learners. The chapter includes a step-by-step tutorial on designing acceptability judgment tasks and truth-value judgment tasks, as well as several study questions that ask readers to design judgment tasks and to interpret sample data.

Schmitt, C., & Miller, K. (2010). Using comprehension methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 35–56). Amsterdam, The Netherlands: John Benjamins.

This book chapter provides an overview of the use of truth-value judgment tasks, picture-matching tasks, and act-out tasks in research with both first and second language learners. The format and administration of each task are discussed, as are its relative advantages and disadvantages. The chapter ends with a list of do's and don'ts in the use of interpretation tasks.

REFERENCES

Abrahamsson, N., & Hyltenstam, K. (2008). The robustness of aptitude effects in near-native second language acquisition. *Studies in Second Language Acquisition*, *30*, 481–509.

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adults' graded judgments of argument-structure overgeneralization errors. *Cognition*, *106*, 87–129.

Bard, E. G., Robertson, D., & Sorace, A. (1996). Magnitude estimation of linguistic acceptability. *Language*, *72*, 32–68.

Bowles, M. A. (2011). Measuring implicit and explicit linguistic knowledge. *Studies in Second Language Acquisition*, *33*, 247–271.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Bruhn de Garavito, J. (2011). Subject/object asymmetries in the grammar of bilingual and monolingual Spanish speakers: Evidence against connectionism. *Linguistic Approaches to Bilingualism*, *1*, 111–148.
Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2010). The role of language aptitude in first language attrition: The case of pre-pubescent attriters. *Applied Linguistics*, *31*, 443–464.
Bylund, E., Abrahamsson, N., & Hyltenstam, K. (2012). Does first language maintenance hamper nativelikeness in a second language? *Studies in Second Language Acquisition*, *34*, 215–241.
Chien, Y.-C., & Wexler, K. (1990). Children's knowledge of locality conditions in binding as evidence for the modularity of syntax and pragmatics. *Language Acquisition*, *1*, 225–295.
Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.
Chomsky, N. (1981). *Lectures on government and binding: The Pisa lectures*. Dordrecht, The Netherlands: Foris.
Clahsen, H., Martzoukou, M., & Stavrakaki, S. (2010). The perfective past tense in Greek as a second language. *Second Language Research*, *26*, 501–525.
Cowart, W. (1997). *Experimental syntax*. Thousand Oaks, CA: Sage.
Crain, S., & Thornton, R. (1998). *Investigations in universal grammar: A guide to experiments on the acquisition of syntax and semantics*. Cambridge, MA: MIT Press.
Dąbrowska, E. (2010). Naïve versus expert intuitions: An empirical study of acceptability judgments. *The Linguistic Review*, *27*, 1–23.
DeKeyser, R. (2000). The robustness of critical period effects in second language acquisition. *Studies in Second Language Acquisition*, *22*, 499–533.
DeKeyser, R. (2013). Age effects in second language learning: Stepping stones towards better understanding. *Language Learning*, *63*, 52–67.
DeKeyser, R., Alfi-Shabtay, R., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics*, *31*, 413–438.
Dekydtspotter, L., Sprouse, R. A., & Anderson, B. (1997). The interpretive interface in L2 acquisition: The process-result distinction in English-French interlanguage grammars. *Language Acquisition*, *6*, 297–332.
Duffield, N., & Matsuo, A. (2009). Native speakers' vs. L2 learners' sensitivity to parallelism in VP-ellipsis. *Studies in Second Language Acquisition*, *31*, 93–123.
Duffield, N., Matsuo, A., & Roberts, L. (2009). Factoring out the parallelism effect in VP-ellipsis: English vs. Dutch contrasts. *Second Language Research*, *25*, 427–467.
Ellis, R. (2005). Measuring implicit and explicit knowledge of a second language. *Studies in Second Language Acquisition*, *27*, 141–172.
Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J., & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing, and teaching*. Bristol, UK: Multilingual Matters.
Falk, Y., & Bardel, C. (2011). Object pronouns in German L3 syntax: Evidence for the L2 status factor. *Second Language Research*, *27*, 59–82.
Fernández, C. (2008). Reexamining the role of explicit information in processing instruction. *Studies in Second Language Acquisition*, *30*, 277–305.
Finer, D. L., & Broselow, E. I. (1986) Second language acquisition of reflexive binding. In S. Berman, J. McDonough, & J.-W. Choe (Eds.), *Proceedings of NELS 16* (pp. 154–168). Amherst, MA: GLSA, University of Massachusetts.
Gabriele, A. (2009). Transfer and transition in the SLA of aspect: A bidirectional study of learners of English and Japanese. *Studies in Second Language Acquisition*, *31*, 371–402.
Gass, S., & Alvarez-Torres, M. (2005). Attention when? An investigation of the ordering effect of input and interaction. *Studies in Second Language Acquisition*, *27*, 1–31.
Geeslin, K. (2003). A comparison of copula choice in advanced and native Spanish. *Language Learning*, *53*, 703–764.
Geeslin, K. (2005). *Crossing disciplinary boundaries to improve the analysis of second language data: A study of copula choice with adjectives in Spanish*. Munich, Germany: LINCOM Europa.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Geeslin, K., & Guijarro-Fuentes, P. (2008). Variation in contemporary Spanish: Linguistic predictors of *estar* in four cases of language contact. *Bilingualism: Language and Cognition*, *11*, 365–380.
Goro, T., & Akiba, S. (2004). Japanese disjunction and the acquisition of positive polarity. In Y. Otsu (Ed.), *Proceedings of the Fifth Tokyo Conference on Psycholinguistics* (137–162). Tokyo, Japan: Hituzi Syobo.
Grüter, T., Lieberman, M., & Gualmini, A. (2010). Acquiring the scope of disjunction and negation in L2: A bidirectional study of learners of Japanese and English. *Language Acquisition*, *17*, 127–154.
Gualmini, A., & Crain, S. (2005). The structure of children's linguistic knowledge. *Linguistic Inquiry*, *36*, 463–474.
Henry, N., Culman, H., & VanPatten, B. (2009). More on the effects of explicit information in instructed SLA. *Studies in Second Language Acquisition*, *31*, 559–575.
Hwang, S. H., & Lardiere, D. (2013). Plural-marking in L2 Korean: A feature-based approach. *Second Language Research*, *29*, 57–86.
Ionin, T. (2012). Formal theory-based methodologies. In A. Mackey & S. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 30–52). Oxford, UK: Wiley-Blackwell.
Ionin, T., & Montrul, S. (2010). The role of L1-transfer in the interpretation of articles with definite plurals in L2-English. *Language Learning*, *60*, 877–925.
Ionin, T., Montrul, S., & Crivos, M. (2013). A bidirectional study on the acquisition of plural noun phrase interpretation in English and Spanish. *Applied Psycholinguistics*, *34*, 483–518.
Ionin, T., Baek, S., Kim, E., Ko, H., & Wexler, K. (2012). *That*'s not so different from *the*: Definite and demonstrative descriptions in second language acquisition. *Second Language Research*, *28*, 68–100.
Ivanov, I. (2012). L2 acquisition of Bulgarian clitic doubling: A test case for the Interface Hypothesis. *Second Language Research*, *28*, 345–368.
Jegerski, J., VanPatten, B., & Keating, G. D. (2011). Cross-linguistic variation and the acquisition of pronominal reference in L2 Spanish. *Second Language Research*, *27*, 481–507.
Jiang, L. (2009). A referential/quantified asymmetry in the second language acquisition of English reflexives by Chinese-speaking learners. *Second Language Research*, *25*, 469–491.
Johnson, J. S., & Newport, E. L. (1989). Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language. *Cognitive Psychology*, *21*, 60–99.
Keating, G., VanPatten, B., & Jegerski, J. (2011). Who was walking on the beach? Anaphora resolution in Spanish heritage speakers and adult second language learners. *Studies in Second Language Acquisition*, *33*, 193–221.
Kim, J.-H., Montrul, S., & Yoon, J. (2009). Binding interpretations of anaphors by Korean heritage speakers. *Language Acquisition*, *16*(1), 3–35.
Kraš, T. (2011). Acquiring the syntactic constraints on auxiliary change under restructuring in L2 Italian: Implications for the Interface Hypothesis. *Linguistic Approaches to Bilingualism*, *1*, 413–438.
Kupisch, T. (2012). Specific and generic subjects in the Italian of German-Italian simultaneous bilinguals and L2 learners. *Bilingualism: Language and Cognition*, *15*, 736–756.
Mackey, A., & Gass, S. (2005). *Second language research: Methodology and design*. Mahwah, NJ: Erlbaum.
Marsden, H. (2009). Distributive quantifier scope in English-Japanese and Korean-Japanese Interlanguage. *Language Acquisition*, *16*, 135–177.
McDaniel, D., McKee, C., & Smith Cairns, H. (Eds.). (1996). *Methods for assessing children's syntax*. Cambridge, MA: MIT Press.
Montrul, S. (1999). Causative errors with unaccusative verbs in L2 Spanish. *Second Language Research*, *15*, 191–219.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

Montrul, S., Dias, R., & Santos, H. (2011). Clitics and object expression in the L3 acquisition of Brazilian Portuguese: Structural similarity matters for transfer. *Second Language Research*, *27*, 21–58.
Murphy, V. (1997). The effect of modality on a grammaticality judgment task. *Second Language Research*, *13*, 34–65.
Papadopoulou, D., Varlokosta, S., Spyropoulos, V., Kaili, H., Prokou, S., & Revithiadou, A. (2011). Case morphology and word order in second language Turkish: Evidence from Greek learners. *Second Language Research*, *27*, 173–205.
Pérez-Leroux, A., Munn, A., Schmitt, C., & Delrish, M. (2004). Learning definite determiners: Genericity and definiteness in English and Spanish. In A. Brugos, L. Micciulla, & C. Smith (Eds.), *BUCLD 28 Online Proceedings Supplement*. Retrieved from http://www.bu.edu/bucld/files/2011/05/28-perez-leroux.pdf
Rebuschat, P. (2013). Measuring implicit and explicit knowledge in second language research. *Language Learning*, *63*, 595–626.
Schmitt, C., & Miller, K. (2010). Using comprehension methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 35–56). Amsterdam, The Netherlands: John Benjamins.
Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics in Language Teaching*, *10*(1–4), 209–232.
Serratrice, L., Sorace, A., Filiaci, F., & Baldo, M. (2009). Bilingual children's sensitivity to specificity and genericity: Evidence from metalinguistic awareness. *Bilingualism: Language and Cognition*, *12*, 239–257.
Slabakova, R., Kempchinsky, P., & Rothman, J. (2012). Clitic-doubled left dislocation and focus fronting in L2 Spanish: A case of successful acquisition at the syntax-discourse interface. *Second Language Research*, *28*, 319–343.
Sorace, A. (2010). Using Magnitude Estimation in developmental linguistic research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 57–72). Amsterdam, The Netherlands: John Benjamins.
Sprouse, J. (2011). A test of the cognitive assumptions of magnitude estimation: Commutativity does not hold for acceptability judgments. *Language*, *87*, 274–288.
Sprouse, J. (2013). Acceptability judgments. In M. Aronoff (Ed.), *Oxford bibliographies online: Linguistics*. Oxford, UK: Oxford University Press.
Stavrakaki, S., & Clahsen, H. (2009). The perfective past tense in Greek child language. *Journal of Child Language*, *36*, 113–142.
Tanenhaus, M., & Carlson, G. N. (1990). Comprehension of deep and surface verb phrase anaphors. *Language and Cognitive Processes*, *5*, 257–280.
Thomas, M. (1991). Universal grammar and the interpretation of reflexives in a second language. *Language*, *67*, 211–239.
Valenzuela, E., Faure, A., Ramírez-Trujillo, A. P., Barski, E., Pangtay, Y., & Diez, A. (2012). Gender and heritage Spanish bilingual grammars: A study of code-mixed determiner phrases and copula constructions. *Hispania*, *95*, 481–494.
VanPatten, B., & Cadierno, T. (1993). Explicit instruction and input processing. *Studies in Second Language Acquisition*, *15*, 225–225.
VanPatten, B., & Borst, S. (2012). The roles of explicit information and grammatical sensitivity in the processing of clitic direct object pronouns and word order in Spanish L2. *Hispania*, *95*, 270–284.
White, L., & Genesee, F. (1996). How native is near-native? The issue of ultimate attainment in adult second language acquisition. *Second Language Research*, *12*, 233–265.
White, L., Hirakawa, M., & Kawasaki, T. (1996). Effect of instruction on second language acquisition of the Japanese long-distance reflexive zibun. *Canadian Journal of Linguistics*, *41*, 135–154.
Year, J., & Gordon, P. (2009). Korean speakers' acquisition of the English ditransitive construction: The role of verb prototype, verb frequency, and input distribution. *The Modern Language Journal*, *93*, 399–417.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press

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

Zhao, L. X. (2012). Interpretation of Chinese overt and null embedded arguments by English-speaking learners. *Second Language Research*, *28*, 169–190.

Zyzik, E. (2014). Causative verbs in the grammar of Spanish heritage speakers. *Linguistic Approaches to Bilingualism*, *4*, 1–33.

Zyzik, E., & Azevedo, C. (2009). Word class distinctions in second language acquisition. *Studies in Second Language Acquisition*, *31*, 1–29.

https://doi.org/10.1017/S0267190514000026 Published online by Cambridge University Press