Chapter 10. When performance masquerades as comprehension: Grammaticality judgments in experiments with non-native speakers

Robyn Orfitelli[1], Maria Polinsky[2]

[1] The University of Sheffield

[2] The University of Maryland

1. Introduction

As experimental work plays an ever-larger role in developing theories of linguistic knowledge, attention to experimental design becomes increasingly important. Sensitivity to methodology is particularly crucial when working with non-native language users. Studies of these groups are less frequently replicated and re-examined than those focusing on the language of child and adult monolinguals, meaning that conclusions based on questionable methodology can have disproportionately far-reaching effects, not only on the theoretical advancement of the field, but also on adult language instruction.

This paper critiques the viability of a particular research methodology in the context of non-native speakers: grammaticality judgment tasks (GJTs).[i] In GJTs, participants are presented with a set of linguistic materials (stimuli) and are either asked whether or not a particular stimulus is “correct” (polar GJT) or to assess the degree to which that stimulus is “correct” (scalar GJT). These elicited judgments constitute the researchers’ data set. GJTs were originally introduced in linguistic research to measure the acceptability of particular language structures for native speakers, the group we will be referring to below as *native controls*. In this capacity, GJTs were used to help diagnose grammatical structure and variation within and across languages. In recent decades, the use of GJTs has expanded to serve as a tool for assessing the grammatical *comprehension* of non-native speakers of a language. Within generative L2 research, the GJT has enjoyed a long history as a metric for determining how native-like a learner’s grammar is across a variety of morphosyntactic phenomena. GJTs are similarly employed with heritage language (HL) speakers. In both cases, it is regularly found that these non-native speakers perform differently from native speakers. Across studies, L2 learners provide a mix of native and non-native judgments in GJTs, while HL speakers consistently perform better on GJTs than early L2 learners, yet still provide non-native judgments. This variable performance is often interpreted as an indicator that L2 and heritage learners operate from a non-native grammatical representation.

The widespread extension of GJTs into L2 research has, by and large, been accepted in the experimental literature, and the results of these tests have featured in some of the fundamental debates in the literature. The present paper draws this practice into question by raising certain red flags about the validity of grammaticality judgments as a measure of comprehension in non-native populations. Broadly, we suggest that poor performance on GJTs by non-native speakers may be related not to grammatical error, but to extra-grammatical factors involving processing and metalinguistic awareness.

We illustrate both the problem and a possible solution by discussing two recent experiments conducted with Russian non-native speakers using GJTs and other tasks. Several pieces are necessary to set up this discussion: Section 2 reviews the use of GJTs in studies of adult L2 and heritage language (HL) speakers. In Section 3, based on numerous instances of within- and across-task inconsistency, we argue that the metalinguistic demands imposed by the task — and the difficulty involved in identifying the root cause of any incorrect answers — render the task unsuitable for testing language comprehension with non-native speakers. Section 4 presents results from two recent Russian HL experiments and uses the data to explore the limits of GJTs in assessing comprehension. Based on these results, in Section 5, we outline several possible alternatives to GJTs that might be embraced in future work with non-native populations, including truth-value judgments and sentence-picture matching tasks. We conclude that (1) GJTs should only be used in studies of non-native populations when no other methodology can appropriately accomplish the same goals, and (2) when GJTs *are* used with non-native participants, the results collected from these studies should be compared against an additional battery of tests to eliminate possible extra-grammatical confounds.

Importantly, the purpose of the present article is not to present a general indictment of the GJT as a linguistic tool for use with native speakers; many valuable insights have been and will continue to be gained through their use. Our critique applies only to cases in which GJTs are used to measure non-native speaker comprehension. Additionally, we intentionally frame our critique in broad terms, acknowledging that some subtypes of GJTs may be less problematic than others when used with non-native speakers.

2. Previous use of GJTs in L2 and HL

We wish to open this section with a brief word on heritage speakers, since this population is still relatively underrepresented in the literature. Heritage language (HL) speakers are bilinguals whose first (home) language does not typically reach native-like proficiency due to a shift (whether abrupt or gradual) to a societally dominant language. Over time, heritage speakers hear and speak the dominant language more than the HL, and by adulthood, are unbalanced bilinguals whose stronger language is no longer the one used by their parents.

Although there is tremendous variation among HL speaker populations, a number of consistent generalizations have emerged which mark adult HL speakers as different from both adult L1 and L2 speakers (Benmamoun et al., 2013; Scontras et al., 2015, and references therein). In general, heritage speakers do well in producing and comprehending simple, unitary structures, but show failures at the discourse level (Laleko, 2010, among others), unlike monolinguals and balanced bilinguals.

In GJT studies, L2 and HL learners both show non-target patterns of responses. The majority of findings suggest that L2 learners perform better on visual GJTs than on auditory ones (e.g., Johnson, 1992; Murphy, 1997, but cf. Abrahamsson & Hyltenstam, 2008, 2009), while the inverse is true for HL speakers, (Montrul et al., 2008), perhaps due to lower literacy levels in their HL. Literacy is not the only cause of non-native judgments, however, as HL speakers also provide non-native judgments on aural GJTs for a range of phenomena (e.g., Knightly et al., 2003; Sherkina-Lieber, 2011).

Although the use of GJTs with these two non-native populations is widespread, it is not without controversy. In an early critique of their use in L2 research, Ellis (1991) presented data from Chinese learners of English who showed great variability on the same GJT conducted twice, one a week apart. Johnson et al. (1996) have observed that L2 learners' responses change when retested on the same task. Notably, in both cases, participants' judgments were more native-like on second testing. Within-subject inconsistencies on GJTs have also been found within the same task. Ellis, for example, reports on numerous subjects changing their judgments on the same task within the same testing period. He remarks:

> The learners showed a marked reluctance to acknowledge their uncertainty, even though the test allowed them to do so. They preferred to use a number of test-taking strategies to arrive at a definite judgment… This might have been because, as experienced test takers, they believed it was always advisable to reach a clear decision (Ellis, 1991, p. 181).

In many GJT studies, L2 learners' correct response rates range between 30 and 60 percent. A 30 percent fluctuation in success rate would be considered highly unusual for native speakers, and it would be surprising for L2 speakers as well if these judgments were based on a consistent (if not-quite-native) underlying grammar.[ii]

L2 learners and HL speakers alike perform worse on GJTs involving ungrammatical items (Ellis, 1991; Flege et al., 1999; Gass, 1983; Johnson et al., 1996; Juffs & Harrington, 1996; Murphy, 1997; Orfitelli

& Grüter 2014; White, 1985, 1986). Both groups correctly then to accept grammatical structures, but are reluctant to reject ungrammatical ones. When L2 learners show improvement at a task upon re-testing, their improvement is always in their rejection of the ungrammatical items (Johnson et al., 1996). Furthermore, even when L2 learners correctly reject ungrammatical items, their reaction times are slower than for grammatical items, an effect that is not seen in native speakers (Murphy, 1997).

Similar effects are found for HL learners. In a large survey of 70 native and 70 heritage speakers of Russian, Polinsky (2006) found that heritage speakers accepted the majority of grammatical sentences *and* many ungrammatical ones in the realms of binding, gender agreement, gerund control, and irregular verbal morphology. For example, heritage speakers rejected only 32 percent of the 100 violations of gender agreement, compared to a 97 percent rejection rate by native speakers. Common responses to ungrammatical conditions from the heritage speakers included "maybe," "I don't know," and other equivocal replies (Polinsky, 2006, pp. 196-199). Sherkina-Lieber (2011) notes that, "[t]he most common error for [higher proficiency heritage speakers] was to accept both the grammatical and ungrammatical sentences in a pair" (Sherkina-Lieber, 2011, p. 181; see also Vishwanath, 2013).

Moreover, replacing binary judgments with rating scales does not resolve L2 and HL participants' hesitancy to reject ungrammatical structures. In a study of Korean native speakers and highly proficient HL and L2 speakers, Laleko & Polinsky (2013, 2015, 2016) found that, on grammatical and marginally acceptable sentences, heritage speakers patterned with native controls in judgments on a 1-5 Likert scale; however, HL speakers' ratings of ungrammatical stimuli tended to be higher than those of native speakers, suggesting that they are generally reluctant to reject ungrammatical items.

How can we explain a selective difficulty with ungrammatical items? In the literature, this pattern is typically attributed to a "yes bias" and largely ignored. However, on the assumption that GJT performance reflects the participant's consistent non-native grammar, it should be surprising that that interlanguage grammar would selectively affect learners' judgments of ungrammatical sentences, but not grammatical sentences.

3. Problems with the GJT and non-native populations: Extra-grammatical confounds

The studies discussed in the previous section reveal a range of patterns in non-native responses to GJTs that are difficult to account for using grammatical principles:

(1) Patterns in the responses

a. Inconsistency in judgments (which cannot be explained in relation to the grammatical feature under investigation), within and across speakers

b. Differences between GJT and production results

c. Differences in performance on grammatical vs. ungrammatical judgments

Taken together, we believe that these response patterns suggest an extra-grammatical explanation for poor GJT performance. If the metalinguistic cognition required by GJTs leads to an increased expenditure of processing resources, non-native speakers' performance may be impeded, even if their underlying grammatical knowledge is native-like with regard to the area under investigation. To take just one example, Tokowicz & McWhinney (2005) found that beginner learners of Spanish who performed very poorly on a GJT of gender and tense showed native-like brain activity in response to violations in the same grammatical domain.

Why should this be? As it turns out, there are good reasons to believe that processing limitations may affect non-native grammaticality judgments. Even for monolingual speakers, there exists a clear link between verbal working memory and grammaticality judgments. When tested on sentences that are difficult to parse, but not strictly ungrammatical (e.g., multiply center-embedded relative clauses), monolinguals with lower verbal working memory scores provide lower grammaticality ratings than those with higher verbal working memory scores (Casasanto et al., 2010). When placed under substantial working-memory strain (through a digit-retention task or similar), native speakers can even be induced to provide incorrect judgments on sentences containing agreement violations, omissions, and even word-order errors (Blackwell & Bates, 1995).

If parsing and working-memory strain can affect native speakers under "exceptional" circumstances, we may expect to observe similar effects in L2 learners even without additional working-memory stressors.[iii] Research by McDonald (2000, 2006) has shown that both native and non-native speakers exhibit selective difficulties in the same areas; there is a direct correlation between higher judgment error rates among L2 learners and slower response times among native speakers in terms of the areas affected (McDonald, 2000).[iv]

The response patterns for non-native speakers discussed in the previous section include inconsistent judgments under task repetition and decreased performance on judgments of ungrammatical sentences. In the case of Ellis's (1991) task repetition quandary, substantial improvement over the course of a week would be unexpected if the participants' initial answers were driven by consistent, underlying grammatical rules; on the other hand, such a pattern is entirely plausible if the GJT is indexing an extra-grammatical skill that can be improved by repetition (either through a lessening of the associated processing demands or through task practice). Similarly, in the realm of ungrammatical sentences, learners faced with a processing difficulty may be inclined to accept sentences they cannot parse in a manner that a non- (or less) processing-impaired learner will not.

An important caveat is necessary here: the identification of a relationship between GJTs and processing ability does not mean that results from GJTs can *never* reflect L2 interlanguage, nor are we asserting that every non-native judgment reported in the literature reflects a processing deficit. What we instead suggest is that the potential confound from extra-grammatical factors makes it extremely difficult to draw conclusions about learners' underlying grammar based on GJT data alone.

4. Eliminating extra-grammatical confounds: Seeking alternatives to the GJT

4.1 GJTs versus comprehension-based tasks

If the GJT is directly affected by processing-related factors — and non-native speakers' judgments are particularly subject to this influence — then GJTs alone cannot be relied on to measure this population's abilities. In this section, we will advocate the use of a suite of *comprehension tests* to measure L2/HL grammatical representations. Native speakers' performance on comprehension tasks has been shown not to be impaired under processing strain, suggesting that unlike GJTs, comprehension tasks are not subject to extra-grammatical confounds (Dick et al., 2001; Waters et al., 1995; Waters et al., 2003). This difference may stem from two sets of factors. First, comprehension tasks do not require metalinguistic reasoning to the same extent that GJTs do; second, participants in comprehension tasks may receive more contextual information, which can facilitate their performance.

If this reasoning is correct and comprehension tests are "safer" than GJTs from a processing perspective, we should expect to find a within-subjects contrast between L2 learners' performance on GJTs and on comprehension-based tasks. Such a prediction is supported by a recent study conducted by Orfitelli & Grüter (2014) on adult Spanish speakers in the early stages of learning English.

Past GJT studies of the "null-subject (NS) parameter" have reported a ~30-40 percent acceptance of ungrammatical NS sentences among native Spanish L2 learners of English (Davies, 1996; White, 1985, 1986). The standard explanation for these data is that Spanish speakers initially transfer their native *pro*-drop grammar into their L2 English. The validity of this explanation is challenged, however, by the fact that, while Spanish learners of English tend to *accept* NS sentences, very rarely do they *produce* these ungrammatical sentences (Lakshmanan, 1994; Phinney, 1987; Ruiz de Zarobe, 1998).

Orfitelli & Grüter (2014) compared results from the standard GJT on NS sentences to results from a truth-value judgment task (TVJT; see Crain & Fodor, 1993). In the TVJT, participants first heard a short story and then heard a sentence in either the imperative or declarative mood. Their task was to judge whether this sentence matched the story they had heard (see Orfitelli & Grüter, 2014, for further details). The experiment hinged on null subject sentences such as: (1) An adult English speaker can assign only an imperative interpretation to this sentence. However, if null subjects undergo transfer from Spanish into L2 English, then learners may accept (2) with a declarative interpretation.[v]

(2) Play with blocks.

The findings from the GJT in this study replicated previous studies: only 55.1 percent of ungrammatical sentences with null subjects were rejected (range: 0-100 percent), with acceptance rates correlating negatively with participant English proficiency ($r(15)=-.80$, $p<.05$). In contrast, when the TVJT results were examined, using a logistic regression with accuracy as the binomial dependent variable, proficiency was not found to be a significant predictor of performance (p=.86). This is because even the lowest-proficiency English learners performed like native English-speaking controls (M=93.38, SD=7.30).[vi]

The discrepancy between the TVJT results and the GJT results suggests that L2 learners' non-native judgments on NS sentences in GJTs are actually caused by extra-grammatical difficulties. These findings suggest a need to reanalyze the observation that acceptance rates decline over time;

rather than indicating a resetting of L1 parameters to L2 settings, this trend could reflect a gradual expansion in the processing resources that L2 learners are able to bring to bear on the GJT.

Orfitelli and Grüter's study strongly suggests that comprehension tasks, such as the TVJT, provide a more reliable measure of L2 interlanguage grammars than the GJT. A study of Inuttitut speakers by Sherkina-Lieber (2011) shows similar results for HL interlanguage grammars. Again in this study, participants took part in both a GJT task and a task measuring comprehension of tense morphemes. In striking contrast to their poor performance on GJTs for items with tense-related violations, the HL speakers gave a native-like performance on the comprehension task, suggesting that they have a native-like representation of tense. Furthermore, none of the production metrics related to the use of agreement and tense was found to correlate with performance on the GJT (Sherkina-Lieber, 2011, ch. 7). Taken together, the contrast between native-like production and comprehension of tense vs. metalinguistic knowledge of tense supports the conclusion that, for HL speakers as for L2 learners, mistakes on GJTs do not arise from fundamental problems in the interlanguage grammar.

4.2 GJT versus comprehension tasks in two Russian HL experiments

To explore further possible comprehension-based replacements for the GJT, in this section we present data from two recent studies of heritage Russian. The first is a within-subjects study of 37 heritage Russian speakers recently conducted by Maria Polinsky. This study compared heritage Russian speakers' performance on a GJT and a sentence-picture matching (SPM) task to the performance of 33 native-speaker controls.

In the GJT, participants were presented with subject relative (SR) clauses where the relative pronoun had either a correct nominative case marking (3) or an incorrect accusative case marking (4). All relative clauses were unambiguous and non-reversible, involving an animate agent and an inanimate theme. The agent and the theme were in opposite genders; since the relative pronoun in Russian indexes gender, there could thus be no confusion as to the relative pronoun's intended referent. In addition, the predicate of the relative clause was placed in the past tense so that it would agree explicitly in gender with the head noun. The repeated indexing of gender was intended to make the stimuli maximally conducive to providing native-like judgments.

(3) a. Ja uvidel starušku [kotor-aja fotografirovala sobor u reki]
1SG saw old_lady.F.ACC which-NOM.F photographed cathedral.M.ACC by river
b. Ja uvidel mal'čika [kotor-yj fotografiroval cerkov' u reki]
1SG saw boy.M.ACC which-NOM.M photographed church.F.ACC by river
'I saw an old lady/boy who was taking pictures of the cathedral/church by the river.'

(4) a. *Ja uvidel starušku [kotor-uju fotografirovala sobor u reki]
1SG saw old_lady.F.ACC which-ACC.F photographed cathedral.M.ACC by river
b. *Ja uvidel mal'čika [kotor-ogo fotografiroval cerkov' u reki]
1SG saw boy. M .ACC which-ACC.M photographed church.F.ACC by river
'I saw an old lady/boy who was taking pictures of the cathedral/church by the river.'

Participants rated 20 pairs of sentences on a 1-5 scale, with 1 being “completely ungrammatical” and 5 being “completely grammatical.” Table 10.1 shows participants’ ratings by stimulus type:

	Grammatical	Ungrammatical
Heritage	4.15 (*SD*=0.36)	3.82 (*SD* =0.46)
Native	4.06 (*SD* =0.34)	2.17 (*SD* =0.33)

Table 10.1 *Average group ratings, Russian relative clauses, 1-5 scale.*

Independent samples t-tests were conducted to compare native controls’ and heritage speakers’ responses. On grammatical sentences, the two groups were not found to differ (*t*(1398) = 1.01, *p*=0.31), but their responses to ungrammatical sentences differed significantly (*t*(1398) = 3.36, *p*<0.001). While native speakers provide low ratings for sentences like (5), heritage speakers’ ratings did not significantly differ from their ratings for the grammatical sentences (*t*(1478) = 0.468, *p*=0.64).

The same subjects were then presented with an auditory sentence-picture matching (SPM) task. Participants were shown two pictures in a random orientation on a screen as they listened to a sentence, and were instructed to click on the picture that best corresponded to the sentence. The SPM task was designed to test speakers’ *interpretation* of a sentence, thereby providing indirect information about their grammatical representations. Test sentences included both subject relative (SR) and object relative (OR) clauses (20 items each). As in the GJT, case marking ensured that all test sentences were unambiguous; however, in the SPM, all referents were inanimate, and as such, the pictures represented reversible situations (e.g., a statue supporting a column vs. a column supporting a statue).

Results of the SPM task found that heritage speakers provide less accurate judgments on ORs than on SRs (Table 10.2), supporting previous findings reported in Polinsky (2011, for L1, L2, and heritage acquisition). For comparison to the average ratings on the GJT, however, it is the SR accuracy data that are most relevant, since these were the structures tested in both tasks.

	Object RC	Subject RC
Heritage	69.02%	90.48%
Native	90.86%	91.53%

Table 10.2 *Accuracy on sentence-picture matching task, Russian unambiguous RCs*

As Table 10.2 shows, HL speakers' comprehension of SRs in the SPM was native-like, suggesting that they have a native-like grammatical representation of these sentences.[vii] This finding distinctly contrasts with their non-native metalinguistic assessments of similar clauses in the GJT.

The findings from this study are supported by empirical results from another recent study of Russian HL speakers conducted by Xiang et al., (2011). This study focused on Russian numerical expressions, a well-known domain of complexity in Russian grammar. Considering only the nominative case, Russian numerical expressions can be divided into three groups based on the morphology of the enumerated noun. With ‘one,’ the noun appears in the (nominative) singular, and the numeral shows gender agreement with the noun. With ‘two’ to ‘four’, the noun appears in the special paucal form, which for most nouns is syncretic with the genitive singular. For ‘five’ and above, the noun is marked as genitive plural, and there is no gender agreement. The following examples illustrate these three patterns for each gender (feminine, masculine, and neuter):

(5) odin-Ø mal’čik-Ø/ odn-a devočk-a/ odno jablok-o
one-MASC boy-NOM.SG/ one-FEM girl-NOM.SG/ one-NEUT apple-NOM.SG
‘one boy, one girl, one apple’

(6) tri mal’čik-a/ tri devočk-i/ tri jablok-a
three boy-GEN.SG/ three girl-GEN.SG/three apple-GEN.SG
‘three boys, three girls, three apples’

(7) šest’ mal’čik-ov/šest’ devoček/ šest’ jablok
six boy-GEN.PL/ six girl.GEN.PL/ six apple.GEN.PL
‘six boys, six girls, six apples’

Xiang et al., (2011) established native speakers’ responses to grammatical numerical phrases like (5)-(7), as well as ungrammatical ones where the features of the numeral and the noun did not match. The mismatches could occur in case form (nominative, paucal, genitive plural), in number (singular, plural), or both. Thus, for this study, each numeral appeared in four contexts, three of which were ungrammatical. Examples of the experimental stimuli are presented in Table 10.3.

W1: PP	W2: Numeral	W3: Adj	W4: Noun (‘boy’)	W5: PP	W6: Predicate	W7: PP
V xore in choir	odin one.NOM	malen’kij small.NOM.SG	mal⍰čik NOM.SG *mal⍰čiki NOM.PL *mal⍰čika PAUCAL *mal⍰čikov GEN.PL	v očkax in glasses	Stojal stood.SG	vperedi in front
	tri three.NOM	malen’kix small.GEN.PL	mal⍰čika PAUCAL *mal⍰čik NOM.SG *mal⍰čiki NOM.PL *mal⍰čikov GEN.PL		Stojali stood.PL	
	pjat’ five.NOM		mal⍰čikov GEN.PL *mal⍰čik NOM.SG		Stojali stood.PL	

			*mal⍰čiki NOM.PL *mal⍰čika PAUCAL			

Table 10.3 *Example stimuli, self-paced reading experiment in Russian (Xiang, Harizanov, Polinsky & Kravtchenko, 2011)*

All participants responded to a GJT (using a 1-7 scale) and to an online self-paced reading task (see Xiang et al., 2011, for experiment details and results for native controls, N=37). The judgments from the native controls were compared with judgments produced by 27 Russian HL speakers for both tasks. (Note that all the HL participants in this study were by necessity relatively proficient and literate in Russian, given the reading requirement of the task.)

For purposes of the present discussion, we report just the comparison of the findings for grammatical and ungrammatical numerical expressions (Figure 10.1). In the off-line rating task (1-7 scale, 1: unacceptable, 7: fully acceptable), a 3 by 4 ANOVA of native-speaker data shows a strong main effect for grammatical vs. ungrammatical sentences (F1(3, 102) = 81, p_1 < .001; F2(3, 177) = 220, p_2 < .001). In contrast, heritage speaker data demonstrate a grammaticality effect only in the context of the numeral 'one' (F1(3, 93) = 82, p_1 < .001; F2 (3, 195) = 321, p_2 < .001). There was no effect of grammaticality on judgments in the other numerical contexts (F(3, 327)=.427; p=.42 for the paucal expressions and F(3, 311)=.832, p=0.34 for the 5+ expressions).

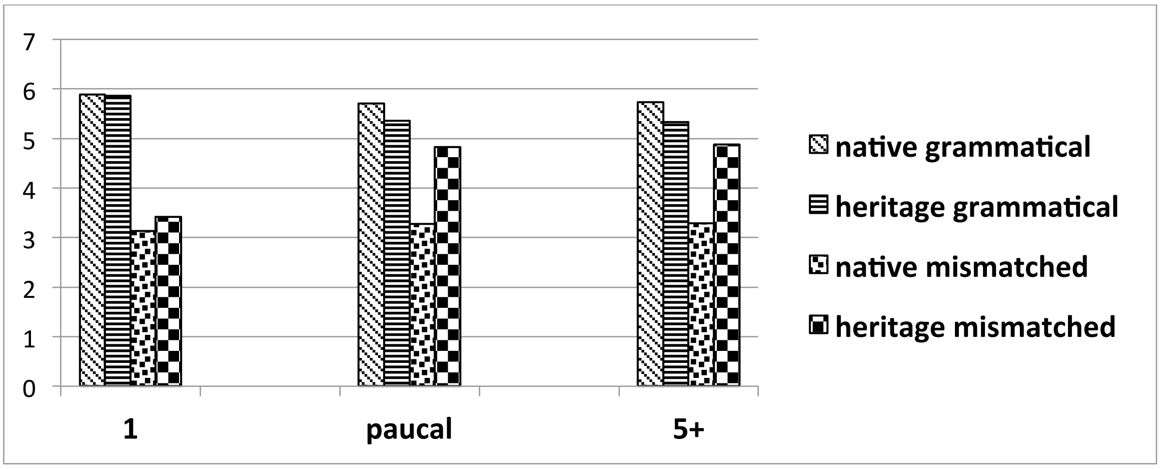

Fig. 10.1. Rating of grammatical and ungrammatical numerical expressions, native and heritage Russian speakers

In contrast to the GJT findings, in the self-paced reading study, both native controls and HLs showed a very similar pattern. Although the HL readers were generally slower than the controls to complete their responses, both groups read the grammatical forms faster than the ungrammatical forms, both at the critical word and at the spillover region (the word[s] following the critical word). As in the original study with native controls, HLs clearly distinguished the fully grammatical and fully ungrammatical forms in the self-paced reading study and read the ‘‘intermediate’’ ungrammatical forms (with partially matching features) exactly like the fully ungrammatical ones. Table 10.4 shows the reading times for both participant groups at the critical noun (following the numeral) and the next word.[viii]

	1	Paucal	5+
	Critical word (N4)		
Grammatical	465 (155)/760 (447)	501 (177)/825 (449)	489 (146)/896 (537)
Case mismatch	540 (219)*/910 (542)*	521 (212)/812 (473)	537 (183)/965 (513)*
Number mismatch	546 (208)**/998 (581)*	507 (180)/857 (432)	538 (198)*/965 (513)*
Case+Number mismatch	576 (233)***/990 (555)**	507 (191)/972 (541)	517 (168)/746 (423)
	Spill over region (W5)		
Grammatical	425 (121)/484 (185)	430 (110)/511 (217)	414 (115)/471 (184)
Case mismatch	496 (140)**/599 (300)*	455 (115)/531 (260)	473 (124)**/499 (224)*
Number mismatch	504 (149)**/593 (224)**	465 (104)/577 (250)	469 (120)*/583 (276)*
Case+Number mismatch	512 (155)***/590 (264)***	441 (109)/590 (290)*	513 (144)***/565 (317)*

Table 10.4 *Residualized reading times (standard derivations) for the critical N and the spillover word, native/heritage speakers*. Significance codes (***.001, **.01, *.05) reflect *p*-values for mixed-effects models with the match-both condition as intercept.

Again, we find that, although Russian HL speakers performed non-natively on the GJT, they performed like Russian native speakers on a comprehension (online self-paced reading) task. These results are reminiscent of the previously mentioned findings by Tokowicz & McWhinney (2005) for L2 learners, and reiterate the overall conclusion presented here, namely, that the GJT may be targeting performance rather than comprehension.

5. Discussion

These two recent studies of Russian heritage language, taken in conjunction with previous L2/HL studies with GJTs discussed in the previous sections, provide compelling evidence that, in at least some areas of the L2 and HL grammar, learners' *comprehension*, as measured by comprehension tasks, exceeds their *performance* on acceptability tasks such as the GJT. This finding should not be taken to negate the results found in metalinguistic tasks; it certainly remains true that, when L2 and HL speakers show non-native behavior on a GJT, it indicates some difference between these non-native populations and native speakers. However, it appears that GJT data may not offer as direct a reflection of non-native speakers' unconscious knowledge of *grammar* as has been previously assumed.

If not grammatical competence, then what do GJTs measure? The position we have taken in this paper is that they measure *language-processing resources*. If this speculation is correct, then the selective difficulties that L2 and HL speakers display in judging ungrammatical sentences could be explained by the additional demand on processing resources required to evaluate ungrammatical structures vs. grammatical ones. This claim will require further investigation in future work.

An alternative explanation for non-native speakers' difficulty with GJTs is that highly conscious, metalinguistic judgments on grammar may be particularly difficult for L2 and HL speakers.[ix] These speakers know that they do not *use* the language the way native speakers do, and perhaps more importantly, they are aware that they do not *know* the language natively. For HL speakers in particular, this conviction will likely have been reinforced by family members and educators, who tend to focus on deficiencies (Grosjean, 1985, 2010; Wong & Fillmore, 2000). The explicit nature of GJTs, in which participants are asked to determine whether the totality of a sentence is correctly formed, bears similarities to examinations in a language classroom, and may thus inspire performance anxiety. In such a situation, it is quite reasonable that the previously discussed "yes bias" might emerge, prompting anxious participants to accept sentences without deeply considering each option.

We suggest that both of these explanations may affect non-native speakers' performance on GJTs — and if this is true, the GJT itself may not be an appropriate test of non-native speakers' underlying grammar. Instead, interpretation-based methods such as sentence-picture matching and truth-value judgment tasks may provide a more reliable way of testing non-native speakers' grammatical knowledge, both because they reduce the amount of conscious consideration required and because they provide a context that can reduce processing demands.

Of course, all this discussion should not be taken to imply that the GJT does not have a place in L2 research. Comprehension tasks are by definition restricted to grammatical sentences, so in cases where knowledge of ungrammaticality is directly in question, the GJT is the only available measure. Where the GJT does continue to be used, however, it is important that researchers keep in mind that non-native-like judgments do not necessarily indicate that participants have a non-native-like grammar.

How can we reduce the metalinguistic requirements of the task in such a situation? One possibility is to incorporate rich scenarios into the GJT, producing an alternative task often referred to as an Acceptability Judgment Task (AJT). For grammatical sentences, the AJT is qualitatively similar to comprehension tasks such as the TVJT, requiring learners to determine how “acceptable” a sentence is in a given context. For ungrammatical sentences, however, which are unacceptable irrespective of context, it is less clear whether the AJT assists learners. For instance, Ionin et al. (2011) used an AJT to probe Russian L1 speakers’ mastery of genericity in their L2 English. Across the majority of conditions in which the target sentences were grammatical in the appropriate context, the L2 learners performed natively; however, in the NP-level genericity condition, 15 of the learners showed over-acceptance of ungrammatical indefinite and/or bare singular NPs. These instances of over-acceptance — particularly of sentences like (8a), which did not match the authors’ hypothesized data patterns — might potentially reflect difficulties with sentence processing. If so, the AJT does not provide a viable alternative to the GJT in terms of alleviating processing load.

Processing difficulties with the AJT can be tested in the future by investigating whether learners who accepted sentences like (8a) score lower on independent working memory or processing tasks than learners who successfully reject these sentences.

(8) a. A brown bear is common in these woods.

b. *Brown bear is common in these woods.

The comprehension-based methods discussed above require extremely careful design in order to ensure that a given grammatical structure will appropriately match the relevant picture or scenario. When such care is taken, however, these tasks can provide a way to assess directly speakers' interpretations of sentence structure. Tests of sentence comprehension based on picture matching, where a subject hears a sentence/word and matches it to a picture, are quite common in aphasiology (Caplan et al., 1997; Wassenaar & Hagoort, 2007, among others), child language acquisition (Weissenbron et al., 1990; Weist, 1991, among others), specific language impairment, second-language learning (Grüter, 2005, and references therein), and the visual-world paradigm (Witzel et al., 2012, and references therein).

In concluding this paper, we wish to echo Ellis's assessment:

> Researchers need to acknowledge that metalingual judgments constitute performance data and be prepared to investigate what kind of performances are involved—much more than has been the case in the past (Ellis, 1991, p. 181).

The GJT has an important place in linguistic research, both historically and today, but it is important that we remember its origins as a tool to identify *possible sentences* in a given language. While the GJT may be extended to serve as an indirect metric of native-speaker competence, it is not a test of comprehension, and if it is used with other groups of learners, we must first verify that the populations under study are able and willing to consider the grammar of their language consciously. In cases where participants may feel anxiety owing to lack of proficiency, the paradigm is compromised — and in the area of comprehension, it becomes entirely untenable. When learner *comprehension* is the subject under question, better-suited tasks are readily available and should be employed instead.

References

1. Abrahamsson, Niclas and Kenneth Hyltenstam. “The robustness of aptitude effects in near-native second language acquisition.” *Studies in Second Language Acquisition* 30 (2008): 481–509.
2. Abrahamsson, Niclas and Kenneth Hyltenstam. “Age of onset and nativelikeness in a second language: Listener perception versus linguistic scrutiny.” *Language Learning* 59 (2009): 249-306.
3. Ammerlaan, Tom. “You get a bit wobbly: Exploring bilingual lexical retrieval processes in the context of first language attrition.” *Australian studies* 9 (1996): 92-102.
4. Benmamoun, Elabbas, Silvina Montrul and Maria Polinsky. “Heritage languages and their speakers: Opportunities and challenges for linguistics.” *Theoretical Linguistics* 46 (2013): 129-181.
5. Bialystok, Ellen and Barry Miller. “The problem of age in second-language acquisition: Influences from language, structure, and task.” *Bilingualism: Language and cognition* 2 (1999): 127-145.

6. Blackwell, Arshavir and Elizabeth Bates. "Inducing agrammatic profiles in normals: Evidence for the selective vulnerability of morphology under cognitive resource limitation." *Journal of Cognitive Neuroscience* 7 (1995): 228–257.
7. Caplan, David, Gloria S. Waters and Nancy Hildebrandt. "Determinants of sentence comprehension in aphasic patients in sentence-picture matching tasks." *Journal of Speech, Language, and Hearing Research 40* (1997): 542–55.
8. Casasanto, Laura S., Philip Hofmeister and Ivan A. Sag. "Understanding acceptability judgments: Additivity and working memory effects." In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society.* Cognitive Science Society: Austin, TX, 2010.
9. Crain, Stephen and Janet Fodor. "Competence and performance in child language." In *Language and cognition: A developmental perspective*, edited by Esther Dromi, 141–171. Norwood, NJ: Ablex, 1993.
10. Davies, William D. "Morphological uniformity and the null subject parameter in adult SLA." *Studies in Second Language Acquisition* 18 (1996): 475-493.
11. Dick, Frederic, Elizabeth Bates, Beveryly Wulfeck, Jennifer A. Utman, Nina Dronkers and Morton A. Gernsbacher. "Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals." *Psychological Review* 108 (2001): 759–788.
12. Ellis, Rod. "Grammaticality judgments and second language acquisition." *Studies in Second Language Acquisition* 13 (1991): 161–186.
13. Flege, James E., Grace H. Yeni-Komshian and Serena Liu. "Age constraints on second-language acquisition." *Journal of Memory and Language* 41 (1999): 78-104.
14. Gass, Susan. "The development of L2 intuitions." *TESOL Quarterly* 17 (1983): 273-291.

15. Grosjean, Fran�ois. “The bilingual as a competent but specific speaker-hearer.” *Journal of Multilingual and Multicultural Development* 6 (1985): 467–477.
16. Grosjean, Fran�ois. *Bilingual: Life and reality*. Cambridge, MA: Harvard University Press, 2010.
17. Grüter, Theres. “Comprehension and production of French object clitics by child second language learners and children with specific language impairment.” *Applied Psycholinguistics* 26 (2005): 363-391.
18. Hakuta, Kenji and Daniel D’Andrea. “Some properties of bilingual maintenance and loss in Mexican background high school students.” *Applied Linguistics* 13 (1992): 72–99.
19. Ionin, Tania, Silvina Montrul, Ji-Hye Kim and Vadim Philippov. “Genericity Distinctions and the Interpretation of Determiners in Second Language Acquisition.” *Language Acquisition* 18 (2011): 242-280.
20. Jessner, Ulrike. “A DST model of multilingualism and the role of metalinguistic awareness.” *The Modern Language Journal*, 92 (2008): 270-283.
21. Johnson, Jacqueline S. “Critical period effects in second language acquisition: The effect of written versus auditory materials on the assessment of grammatical competence.” *Language Learning,* 42 (1992): 217–248.
22. Johnson, Jacqueline S. and Elissa L. Newport. “Critical period effects in second language learning: The influence of maturational state on the acquisition of English as a second language.” *Cognitive Psychology* 21 (1989): 60–99.
23. Johnson, Jacqueline S., Kenneth D. Shenkman, Elissa L. Newport and Douglas L. Medin. “Indeterminacy in the grammar of adult language learners.” *Journal of Memory and Language*, 35 (1996): 335-352.
24. Juffs, Alan and Michael Harrington. “Parsing effects in second language sentence processing.” *Studies in Second Language Acquisition*, 17 (1995): 483–516.

25. Juffs, Alan and Michael Harrington. "Garden path sentences and error data in second language sentence processing." *Language Learning,* 46 (1996): 283–326.

26. Knightly, Leah M., Sun-Ah Jun, Janet S. Oh, and Terry K-F. Au. "Production benefits of childhood overhearing." *The Journal of the Acoustical Society of America* 114 (2003): 465.

27. Lakshmanan, Usha. *Universal grammar in child second language acquisition.* Amsterdam: Benjamins, 1994.

28. Laleko, Oksana. "The syntax-pragmatics interface in language loss: Covert restructuring of aspect in Heritage Russian." PhD Diss., University of Minnesota, 2010.

29. Laleko, Oksana, and Maria Polinsky. "Marking topic or marking case? A comparative investigation of Heritage Japanese and Heritage Korean." *Heritage Language Journal,*10 (2013): 40-64.

30. Laleko, Oksana, and Maria Polinsky "Topic-subject asymmetry in Japanese and Korean: Heritage and L2 Speakers." *Harvard Working Papers in Linguistics* 13 (2015): 55-68.

31. Laleko, Oksana, and Maria Polinsky. "Between syntax and discourse: Topic and case marking in heritage speakers and L2 learners of Japanese and Korean." *Linguistic Approaches to Bilingualism* 6 (2016): 396-439.

32. Mayberry, Rachel I. and Elizabeth Lock. "Age constraints on first versus second language acquisition: Evidence for linguistic plasticity and epigenesis." *Brain and Language* 87 (2003): 369–384.

33. Mayo, Lynn H., Mary Florentine and Søren Buus. "Age of second language acquisition and perception of speech in noise." *Journal of Speech Language and Hearing Research* 40 (1997): 686–693.

34. McDonald, Janet L. "Grammaticality judgments in a second language: Influences of age of acquisition and native language." *Applied Psycholinguistics* 21 (2000): 395–423.

35. McDonald, Janet L. “Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners.” *Journal of Memory and Language* 55 (2006): 381–401.

36. McElree, Brian., Gisael Jia and Annie Litvak. “The time course of conceptual processing in three bilingual populations.” *Journal of Memory and Language* 42 (2000): 229–254.

37. Meador, Diane, James E. Flege and Ian R.A. MacKay. “Factors affecting the recognition of words in a second language.” *Bilingualism: Language & Cognition* 3 (2000) 55–67.

38. Montrul, Silvina. *Incomplete acquisition in bilingualism: Re-examining the age factor (Vol. 39).* John Benjamins Publishing Company, 2008.

39. Montrul, Silvina, Rebecca Foote and Silvia Perpiñán. “Gender agreement in adult second language learners and Spanish heritage speakers: The effects of age and context of acquisition.” *Language Learning* 58 (2008): 503–553.

40. Murphy, Victoria A. “The effect of modality on a grammaticality judgment task.” *Second Language Research*, 13 (1997): 34–65.

41. Orfitelli, Robyn and Nina Hyams. “Children's grammar of null subjects: Evidence from comprehension.” *Linguistic Inquiry,* 43 (2012): 563–590.

42. Orfitelli, Robyn and Theres Grüter. “Do null subjects really transfer? In *Proceedings of the 12th Generative Approaches to Second Language Acquisition Conference,* edited by Jennifer Cabrelli-Amano, Tiffany Judy, and Diego Pascual y Cabo. Somerville, MA: Cascadilla Press. 2014.

43. Phinney, Marianne. “The Pro-drop Parameter in Second Language Acquisition.” In *Parameter Setting,* edited by Thomas Roeper and Edwin Williams, 221–238. Dordrecht: Reidel, 1987.

44. Polinsky, Maria. “Incomplete acquisition: American Russian.” *Journal of Slavic Linguistics,* 14 (2006): 191–262.

45. Polinsky, Maria. “Gender under incomplete acquisition: Heritage speakers’ knowledge of noun categorization.” *Heritage Language Journal*, 6 (2008): 40-71.

46. Polinsky, Maria. “Reanalysis in adult heritage language.” *Studies in Second Language Acquisition,* 33 (2011): 305–328.

47. Ruiz de Zarobe, Yolanda. “El parámetro pro-drop en la adquisición de segundas lenguas.//The pro-drop parameter and second language acquisition.” *ITL: International Journal of Applied Linguistics* 119-120 (1998): 49–64.

48. Sanders, Lisa D., Helen J. Neville and Marty G. Woldorff. “Speech segmentation by native and non-native speakers: The use of lexical, syntactic, and stress-pattern cues.” *Journal of Speech Language and Hearing Research* 45 (2002): 519–530.

49. Scherag, André, Lisa Demuth, Frank Rösler, Helen J. Neville and Brigitte Röder. “The effects of late acquisition of L2 and the consequences of immigration on L1 for semantic and morpho-syntactic language aspects.” *Cognition* 93 (2004): B97–108.

50. Scontras, Gregory, Zuzanna Fuchs and Maria Polinsky. “Heritage language and linguistic theory.” *Frontiers in Psychology* 6, 1545.

51. http://journal.frontiersin.org/article/10.3389/fpsyg.2015.01545/full

52. Sherkina-Lieber, Marina. “Comprehension of Labrador Inuttitut functional morphology by receptive bilinguals.” PhD Diss., University of Toronto, 2011.

53. Tokowicz, Natasha and Brian McWhinney. “Explicit and implicit measures of sensitivity to violations in second language grammar.” *Studies in Second Language Acquisition* 27 (2005): 173-204.

54. Vishwanath, Arun. “Heritage English in Israeli children.” BA Thesis, Harvard University, 2013.

55. Wassenaar, Marlies and Peter Hagoort. “Thematic role assignment in patients with Broca's aphasia: Sentence picture matching electrified.” *Neuropsychologia* 45 (2007): 716–740.

56. Waters, Gloria S., David Caplan and Elizabeth Rochon. “Processing capacity and sentence comprehension in patients with Alzheimer’s disease.” *Cognitive Neuropsychology* 12 (1995): 1–30.

57. Waters, Gloria S., David Caplan and Sasha Yampolsky. “On-line syntactic processing under concurrent memory load.” *Psychonomic Bulletin and Review* 10 (2003): 88–95.

58. Weissenbron, Jürgen, Michèle Kail and Angela Friederici. “Language-particular or language-independent factors in acquisition? Children's comprehension of object pronouns in Dutch, French and German.” *First Language* 10 (1990): 141–166.

59. Weist, Richard M. “Spatial and temporal location in child language.” *First Language* 11 (1991): 253–267.

60. White, Lydia. “The “Pro-Drop” parameter in adult second language acquisition.” *Language Learning,* 35 (1985): 47–62.

61. White, Lydia. “Implications of parametric variation for adult second language acquisition: an investigation of the pro-drop parameter.” In *Experimental approaches to second language acquisition*, edited by Vivian Cook, 55–72. Oxford: Pergamon, 1986.

62. Witzel, Naoko, Jeffery Witzel, and Kenneth I. Forster. “Comparisons of online reading paradigms: Eye tracking, moving-window, and maze.” *Journal of Psycholinguistic Research* 41 (2012): 105–128.

63. Wong Fillmore, Lily. “Loss of family languages: Should educators be concerned?” *Theory into Practice* 39 (2000): 203–210.

64. Xiang, Ming, Boris Harizanov, Maria Polinsky, and Ekaterina Kravtchenko. “Processing morphological ambiguity: An experimental investigation of Russian numerical phrases.” *Lingua* 121 (2011): 548-560.

[i] Also referred to as acceptability judgment tasks (AJTs). For simplicity, we use the term GJT to refer to all tasks that elicit metalinguistic judgments about language data, although we specifically touch on the use of context-rich AJTs in the paper's conclusion.

[ii] Within- and across-speaker variability is found with native speakers in tasks that require interpreting subtle semantic contexts or which induce a processing strain. The latter observation further highlights the link between processing resources and performance on GJTs.

[iii] In comparison to native speakers, L2 learners have fewer language-processing resources along several dimensions: speech-decoding ability (Sanders, Neville, & Woldorff, 2002; Mayo, Florentine, & Buus, 1997; Meador, Flege, & MacKay, 2000); effects of semantic priming (McElree, Jia, & Litvak, 2000); and lexical-decision times (Scherag, Demuth, Rösler, Neville, & Röder, 2004). Even highly proficient L2 learners have slower reaction times on GJTs than native speakers, even when they answer the items correctly. This, in turn, suggests slower language processing on the part of L2 speakers overall (Bialystok & Miller, 1999; Mayberry & Lock, 2003; McDonald, 2000; Murphy, 1997; Juffs & Harrington 1995, 1996).

[iv] The study also reports a correlation between GJT performance and age of acquisition (AOA), a relationship which has been extensively documented for L2 learners (see Johnson & Newport, 1989, among others). A similar relationship has been noted for HL speakers and the age of interruption (AOI; see Ammerlaan, 1996; Hakuta & D'Andrea, 1992; Montrul, 2008). Without further study, it is difficult to disentangle the relationship between AOA/AOI and native-like grammatical competence and processing abilities. We believe it is reasonable to assume that AOA/AOI are related to both.

[v] One might expect the imperative prosody assigned to (2) to cue learners to its syntax. However, Orfitelli & Hyams (2012) did not find this prosodic cue to be sufficient for children, who accepted (1) as declarative as well as imperative.

[vi] See Orfitelli & Grüter (2013, nt 4), for specific details of a model run using the lme4 package in R.

[vii] A possible alternative explanation for these data is that Russian HL speakers' grammars permit nouns to be ambiguous in terms of gender. However, evidence from Polinsky (2008) shows that heritage speakers of Russian unambiguously assign each noun to a gender class, even if that class does not match the baseline.

[viii] For all other regions, mixed-effect models revealed no predictive power for either a fixed effect (numerical context and morphological feature matching on the N) or for an interaction between the two. Nor did planned pair comparisons reveal any significant difference between conditions.

[ix] There is evidence (e.g., Jessner, 2008, and references therein) that multilinguals are in fact more metalinguistically aware than monolinguals. The multilinguals on whom this evidence is based, however, are all fairly balanced in their exposure to at least two languages, in contrast to adult HL speakers, whose languages are markedly unbalanced. Furthermore, the metalinguistic advantages discussed for multilinguals arise predominantly in the area of word selection, communicative sensitivity, and translation. As far as we know, increased metalinguistic awareness in the area of grammar has not been noted.