languages

MDPI

Article

Sensitivity to Filler–Gap Dependency Violations in the L1 vs. L2: Evidence from Speeded Judgement Tasks

Aleksandra Trifonova * and Claudia Felser

Potsdam Research Institute for Multilingualism, University of Potsdam, 14476 Potsdam, Germany; claudia.felser@uni-potsdam.de
* Correspondence: aleksandra.trifonova@uni-potsdam.de

Abstract: We carried out four timed judgement experiments investigating whether bilingual speakers differ in their sensitivity to different kinds of filler–gap dependency violation in L1 German and L2 English. Using a within-subjects design and parallel experimental designs for both languages, we manipulated either the availability of a gap ("filled-gap paradigm") or the semantic congruency between the filler and its licensing verb. We examined whether participants exhibited consistent judgement patterns for syntactic (i.e., filled *wh*-gaps) and semantic (i.e., implausible *wh*-fillers) violations within and across their languages. Our results showed that participants' sensitivity to filled gaps correlated positively with their sensitivity to a filler's semantic fit in their L1 but not in their L2, and that participants' sensitivity to semantic fit was positively correlated in their two languages whilst their sensitivity to gap availability was not. Further analyses of the L2 data showed that participants' sensitivity to semantic fit but not to filled gaps increased with L2 proficiency. Our findings are in line with earlier findings indicating reduced sensitivity to structural gaps even at advanced L2 proficiency levels. They also highlight the need for L2 processing research to look beyond group-level performance and consider bilinguals' sensitivity to different types of linguistic constraints at the individual level.

Keywords: L2 processing; filler–gap dependencies; acceptability judgements; German; English

Academic Editors: Line Burholt Kristensen and Sabine Gosselke-Berthelsen

Received: 13 August 2024
Revised: 4 November 2024
Accepted: 12 January 2025
Published: 24 January 2025

Citation: Trifonova, A., & Felser, C. (2025). Sensitivity to Filler–Gap Dependency Violations in the L1 vs. L2: Evidence from Speeded Judgement Tasks. *Languages*, *10*(2), 21. https://doi.org/10.3390/languages10020021

1. Introduction

Syntactically displaced constituents present a well-known challenge for left-to-right incremental sentence processing and have been argued to present a particular challenge for non-native (L2) comprehenders. In *wh*-movement languages like English, a fronted *wh*-phrase such as *what book* in the sentence *What book are you reading?* cannot be fully interpreted locally but must be kept in memory until it can be linked to its lexical licenser (the verb *read*). Only then will it become clear to comprehenders that *what book* functions as the direct object of *read* and carries the thematic role of theme. According to the Active Filler Hypothesis (Frazier, 1987), encountering a fronted constituent (or "filler") will lead the parser to actively search for a corresponding "gap", that is, for an unfilled syntactic position to link the filler to so as to allow for it to be fully interpreted. Establishing a filler–gap dependency (FGD) between *what book* and the "missing" direct object of *read* in *What book are you reading* __? will result in the filler being assigned the verb's theme role and being integrated into the emerging event representation. Syntactic gap-filling and lexically driven filler integration have been shown to be empirically dissociable components of FGD resolution (Felser & Jessen, 2020; Nicol, 1993). While the need for resolving FGDs indirectly via structural gaps has been called into question (Pickering & Barry, 1991), integrating a filler semantically with its lexical licenser is crucial for successful sentence interpretation.

Maintaining a filler in memory while processing subsequent words and phrases, identifying a suitable gap or lexical licenser further downstream, and retrieving a faithful representation of the filler at that point require efficient processing routines and sufficient cognitive resources. This should make FGD resolution particularly challenging for comprehenders whose processing resources are limited or whose processing routines are insufficiently automatized.

Real-time sentence comprehension studies have reported cross-population differences in FGD resolution. Unlike native (L1) speakers, L2 speakers have often failed to show evidence of postulating structural gaps (Berghoff, 2022; Felser & Roberts, 2007; Marinis et al., 2005; Miller, 2015). Felser and Roberts' (2007) results from a cross-modal priming experiment on English indirect object *wh*-dependencies, for example, showed that unlike native English speakers, L1 Greek/L2 English speakers did not retrieve the fronted constituent at the corresponding gap position. Detecting semantic or pragmatic mismatches between a filler and its lexical licenser seems to be less problematic in L2 processing tasks (e.g., Felser et al., 2012; Jessen & Felser, 2019; Williams et al., 2001), although syntactic gap-filling and lexically driven filler integration have not usually been compared in a single study.

Within-population differences in FGD resolution have also been observed. Nakano et al. (2002), for example, found that L1 Japanese comprehenders with a high working memory (WM) span, but not those with a low WM span, retrieved distant fillers at structural gap sites. Berghoff (2023) reports similar WM effects for advanced L1 Afrikaans/L2 English speakers, using a cross-modal picture priming task to investigate indirect object dependency processing in the L1 and L2 (in a close replication of Felser and Roberts' (2007) study).

In neither L1 nor L2 processing does the apparent failure to link a filler to a syntactic gap seem to preclude successful comprehension, however. These observations indicate that FGD resolution does not necessarily require the computation of syntactic gaps but can also be achieved by directly associating a filler semantically with its lexical licenser (Pickering & Barry, 1991).

The above findings raise the possibility that individuals may differ in whether their FGD resolution is primarily gap-driven or semantically driven and that any such individual processing biases are determined by multiple factors, including language status (L1 vs. L2). Most previous L2 processing studies have used between-group designs, which may mask individual variation and make it impossible to relate individual differences in L1 processing to L2 performance. The current study used a violation paradigm to identify bilingual comprehenders' FGD resolution abilities or biases in both of their languages and to examine whether these were shared between the L1 and L2. Testing the same individuals in both of their languages allowed us to examine the effects of language status and L2 proficiency while controlling for individual difference variables (such as WM capacity) that are independent of language status and not a focus of interest here. Our study contributes to the growing body of research on individual differences in L1 and L2 comprehension. Unlike many previous L2 processing studies, we focused on L1 processing performance as a predictor for L2 performance rather than on the potential effects of individual cognitive variables on L2 performance.

1.1. Syntactic and Semantic Filler Integration

In verb-initial languages like English, effects of structural gap-filling and semantic filler integration cannot always be teased apart. One way of empirically dissociating gap-driven and lexically driven filler integration effects is to examine indirect object dependencies. Nicol (1993) used the cross-modal priming technique to examine L1 English speakers' resolution of indirect object *wh*-dependencies as in (1).

(1) To which butcher did the woman who had just inherited a large sum of money give the very expensive gift __ the other day?

Nicol (1993) observed that her participants mentally reactivated the fronted *wh*-phrase *to which butcher* both at the main verb (*give*) and the canonical indirect object position following the direct object (*the very expensive gift*), as indicated by statistically significant priming effects at these two points within the sentence. This suggests that FGD resolution is a two-step process that involves both semantic and syntactic filler integration. Nicol proposed that a verb's argument structure and subcategorization frame are accessible as soon as the verb is encountered and that an initial thematic evaluation of potential argument phrases takes place at this point. The noun phrase (NP) following the verb will be assigned the verb's theme role, and an empty syntactic position will be posited for the missing indirect object in accordance with the verb's subcategorization frame. The fronted prepositional phrase (PP) *to which butcher* is reactivated and linked to this position, allowing for the PP's tentatively assigned goal role to be confirmed.

Besides the cross-modal priming technique, FGD resolution studies frequently use two experimental paradigms to examine filler integration effects. One of these is the so-called "filled-gap" paradigm, where a putative gap is filled by another constituent, giving rise to local or global (Hestvik et al., 2007) ungrammaticality.

(2) *The zebra that the hippo kissed the camel on the nose ran far away.

Filled-gap effects may be reflected, for example, in elevated reading times compared to a control condition without a filled gap (Stowe, 1986) or in participants' brain responses (Hestvik et al., 2007). Filled-gap effects indicate that comprehenders tried to link a filler (here, *the zebra*) to a putative gap and experienced a disruption in processing when finding the gap already occupied. Note that inserting an unlicensed argument phrase such as *the camel* in (2) into the verb's direct object position might be perceived either as a phrase structure violation or as a violation of the verb's argument structure, or possibly both.

A second popular diagnostic of filler integration processes is the manipulation of a filler's semantic fit with its potential licenser (Garnsey et al., 1989).

(3) The businessman knew {which customer/#which article} the secretary called __ at home.

In the semantically infelicitous version of example (3), the *wh*-phrase *which article* functions as the direct object of the verb *called* but fails to meet the verb's selectional restrictions. For sentences containing semantically implausible fillers, increased processing difficulty is observed when the filler's lexical licenser is encountered, relative to sentences containing plausible fillers (such as *which customer* in (3)).

While filled-gap effects might primarily reflect the attempt to integrate a filler into the emerging syntactic representation, filler plausibility effects reflect comprehenders' difficulty in trying to integrate the filler with its licensing verb. Felser and Jessen (2020) used both experimental paradigms to examine and compare L1 German speakers' brain responses to filled-gap and plausibility violations. The authors observed qualitatively different brain responses to filled gaps and implausible fillers, confirming the assumption that dissociable syntactic and semantic integration processes are involved in FGD resolution.

Successful filler integration requires that a faithful representation of the filler is kept in memory until a potential gap or lexical licenser is encountered. It is conceivable that filler representations decay over time or that only partial information about the filler is maintained throughout the dependency length. Wagers and Phillips (2014) examined filled-gap violations, semantic plausibility violations, and subcategorization violations in sentences with varying dependency lengths in three self-paced reading experiments. While the authors observed filled-gap effects at all dependency lengths, filler plausibility effects were substantially delayed in the long dependency condition. According to Wagers and

Phillips' (2014) hybrid maintenance–retrieval model, only coarse-grained information such as a filler's syntactic category is actively maintained over longer distances, enabling the active postulation of syntactic gaps. In contrast, semantic information must be retrieved from memory, which becomes harder with an increasing dependency length.

The authors attribute this difference to our limited focal attention capacity, which only allows for essential information to be kept in focal attention or 'activated' during processing. Information that is deemed non-essential by the parser will be moved to non-focal attention. Since category information is functionally more useful for structure building and syntactic dependency formation than finer grained semantic features, retaining syntactic (but not necessarily semantic) filler information in focal attention is considered vital for gap prediction and gap-filling.

1.2. L2 Comprehension of Filler–Gap Dependencies

Encountering a displaced constituent triggers an active search mechanism in both L1 and L2 comprehension (Al-Maani et al., 2024; Boxell & Felser, 2017; Felser et al., 2012; Omaki & Schulz, 2011; Williams et al., 2001). L2 comprehenders do not consistently show nativelike FGD resolution patterns, however. Studies using the cross-modal priming paradigm have failed to find evidence for position-specific filler retrieval in L2 auditory comprehension (Felser & Roberts, 2007; Miller, 2015; but cf. Berghoff, 2023), and reading-time studies have found evidence for the postulation of structurally defined gaps in L1 but not in L2 comprehension (Berghoff, 2022; Marinis et al., 2005; but cf. Pliatsikas & Marinis, 2013).

Studies investigating L2 speakers' sensitivity to filled gaps have reported mixed findings. Results from self-paced reading tasks have indicated immediate sensitivity to filled gaps in L2 processing (e.g., Al-Maani et al., 2024; Aldwayan et al., 2010; Johnson et al., 2016). Al-Maani et al. (2024), for example, presented L2 English speakers with sentences such as (4), where the first potential licenser of the interrogative pronoun *who* is the verb *put*, whose object position however turns out to be filled by the proper name *Liz*.

(4) My cousin wondered who David will put Liz near __ at the wedding.

The authors report that both L1 Arabic and L1 Mandarin speakers showed increased reading times at filled object gaps in comparison to a non-gap control condition. In an eye-movement monitoring study investigating FGD processing in English, Felser et al. (2012) found filled-gap effects to be delayed in L1 German/L2 English speakers relative to an L1 English-speaking control group, however.

Recording L1 and L2 English speakers' brain responses to filled gaps in sentences such as (2) above, Dong et al. (2023) found that the L1 and L2 groups showed different brain responses to superfluous argument phrases. While the L1 group showed the expected "P600" response typically elicited by syntactic violations, the L2 group showed a prefrontal–central positivity, which the authors suggest may reflect meaning-based rather than structure-based processing. Recall that in sentences such as (2), the filled-gap violation could potentially be perceived either as a syntactic or semantic anomaly.

Jessen et al. (2017) measured brain responses during L1 and L2 speakers' processing of sentences containing filled indirect object gaps, such as (5).

(5) *Sarah tickled the monkey for which Peter arranged some classes for it after the vacation.

Unlike Dong et al. (2023), the authors found that encountering a filled gap triggered a P600 brain response in both participant groups. Note that by using filled gaps that can be interpreted as resumptive (*for which . . . for it*), the authors tried to reduce the possibility of the violation being perceived as a semantic rather than a syntactic one. The P600 effect was stronger and more widespread in the L2 group, which indicates that encountering a filled gap caused a greater processing disruption in L2 than in L1 comprehension. Covey

et al. (2024) also observed a P600 response to (non-resumptive) filled gaps in L2 English, which the authors argue reflects syntactic integration difficulty. The P600 effect was absent in Covey et al.'s (2024) L1 control group, however, who showed a qualitatively different brain response to filled gaps.

More consistent evidence of nativelike L2 processing has been obtained for the detection of semantic mismatches between a filler and its lexical licenser. Williams et al. (2001), for example, found that L2 English speakers from typologically different L1 backgrounds showed immediate sensitivity to implausible fillers in a reading-time study. Jessen and Felser (2019) observed that implausible fillers elicited the same brain responses in both L1 and L2 English speakers when a potential lexical licenser was encountered. In their eye-movement monitoring study, Felser et al. (2012) found that L1 German/L2 English speakers reacted even earlier to implausible fillers than L1 English speakers did.

One exception to these rather consistent findings are the results reported by Dallas et al. (2013). Recording L1 and L2 English speakers' brain responses to filler plausibility violations as in example (6), the authors observed no plausibility effect at the licensing verb (here, *threatened*) in the L2 group.

(6) The umpire asked {which player/#which football} the coach threatened __ before the game.

Additional analyses revealed that the likelihood of L2 speakers showing sensitivity to semantic incongruency increased with L2 proficiency. Besides being influenced by proficiency (Al-Maani et al., 2024; Dallas et al., 2013), FGD resolution in an L2 has been found to be modulated by the type of task (Williams, 2006) and by individual difference variables such as L2 exposure (Berghoff, 2023), WM capacity (Berghoff, 2022), or attentional control abilities (Covey et al., 2024).

The results from previous L2 FGD resolution studies are difficult to compare because they have used different methods, different types of experimental stimuli, and L2 speakers from varied L1 backgrounds and with different L2 acquisition histories. Previous studies have also typically used between-group designs. However, comparing groups of different individuals carries the risk of the results being affected by individual differences in cognitive or other non-linguistic variables or by individual processing abilities or biases.

1.3. Accounting for Variability in FGD Resolution

On the assumption that L2 comprehension is cognitively more demanding than L1 comprehension, processing resource limitations are a possible reason why L2 speakers sometimes fail to resolve FGDs in a nativelike way. According to the Processing Difficulties Hypothesis (McDonald, 2006), resource limitations should have a particularly strong impact on the processing of complex structures. However, the hypothesis that FGD resolution may overburden L2 speakers' processing system cannot account for the observation that L2 comprehenders are more likely to show difficulty with structural gap processing than with semantically driven filler integration.

Experimental findings showing that L1 but not necessarily L2 speakers postulate structural gaps during the processing of FGDs were part of the motivation underlying the development of the Shallow Structure Hypothesis (SSH; Clahsen & Felser, 2006; 2018). This hypothesis states that L2 comprehenders are less likely than L1 comprehenders to compute fully detailed grammatical representations during processing, whilst their ability to use lexical–semantic cues to sentence interpretation is assumed to be good. Weighted constraint-based (Smolensky et al., 2014) or cue-based (Lewis & Vasishth, 2005; MacWhinney, 2008) processing models are suitable for capturing this hypothesized L1/L2 difference. We might assume the relative weighting of structural cues to interpretation to be reduced in L2 compared to L1 comprehension, or the weighting of non-structural cues

or constraints to be increased in L2 relative to L1 comprehension, or both (Clahsen & Felser, 2018; Cunnings, 2017).

Variability in FGD resolution may not be limited to cross-population differences, however. As noted above, traditional group-level analyses can mask within-group inter-individual variability in participants' processing ability or cue sensitivity. Within-group variability in grammatical processing has been observed both among L1 and L2 speakers. While there has been considerable interest in the role of individual difference factors in L2 processing and acquisition (Kidd et al., 2018), few studies have examined the extent to which an individual's semantic and structural processing abilities are correlated (Hopp, 2015) or whether semantic and/or syntactic processing abilities are correlated across an individual's L1 and L2. A notable exception is a study by Grey (2023), who recorded L1 English/L2 Spanish speakers' brain responses to grammatically or semantically anomalous sentences in both English and Spanish (see example 7 for illustration).

(7) The sea lions can {bask/*basking/#edit} on the beach all day.

Grey observed individual variability in participants' brain responses in both L1 and L2 processing and in participants' brain responses to both syntactic and semantic processing. She interpreted her results as evidence of different processing routes to comprehension, with some individuals using a 'semantic' route and some a 'syntactic' route to comprehension (Tanner et al., 2013). Individual participants' brain responses to grammatical and semantic violations were correlated in their L1 but not in their L2, and their brain responses were not correlated between their two languages.

As noted by Grey (2023) and others, evidence of what appears to be qualitative variability in both L1 and L2 comprehension calls into question the legitimacy and usefulness of contrasting "nativelike" and "nonnativelike" processing performance. To capture individual differences and similarities between L1 and L2 processing, the hypothesis that L1/L2 processing differences are gradient (Clahsen & Felser, 2018; Cunnings, 2017) may need to be extended to the individual level (Yadav et al., 2022).

Assuming that FGD resolution involves dissociable syntactic and semantic processes, filler–gap phenomena offer another possibility for investigating whether individuals have different processing profiles. It is conceivable that language status (L1 vs. L2) and individual processing biases interact in some way that previous studies were not designed to investigate. That is, individual comprehenders may be good or bad at resolving FGDs or may be biased towards either gap-driven or semantically driven FGD resolution, independently of whether the language under investigation is their L1 or an L2. The present study used a within-group design to systematically explore this possibility.

2. The Present Study

While a substantial body of research has investigated whether syntactic representations are shared between bilingual speakers' languages (Hartsuiker et al., 2004), the question of whether individual processing abilities or biases are shared across a bilingual person's languages has rarely been examined (Grey, 2023). Here, we used timed judgement tasks to investigate the effects of language status on FGD resolution ability in L1 German/L2 English speakers. We asked whether the same individuals differed in their sensitivity to different kinds of FGD violations in their L1 and L2 and examined whether participants exhibited consistent judgement patterns for syntactic and semantic violations within and across their languages. As both German and English are *wh*-movement languages, any potential difficulty resolving FGDs in the L2 could not be due to our participants' lack of familiarity with constituent fronting. The fact that both languages also share the same script should preclude the possibility of participants' processing being hampered by orthographic decoding difficulty.

We carried out four binary judgement experiments, two in German and two in English. For each language, one experiment probed participants' sensitivity to filled *wh*-gaps, while the other one probed their sensitivity to implausible *wh*-fillers. We used speeded word-by-word stimulus presentation and timed judgements so as to induce participants to process our stimuli incrementally and to reduce the possibility of participants drawing on their metalinguistic knowledge when providing their judgements (Blackwell et al., 1996).

Filled-gap (FG) violations were created by inserting a superfluous pronominal constituent into object gap positions, and filler plausibility (FP) violations by using object fillers that failed to meet the verb's selectional restrictions. Our study design allowed us to address the following specific research questions:

(i) How does language status (L1 vs. L2) affect our bilingual group's sensitivity to filled-gap and filler plausibility violations?
(ii) Are individual participants' sensitivity scores for different types of FGD violation correlated within and/or across their two languages?

Regarding possible effects of language status, we tested the following predictions. From the perspective of Wagers and Phillips' (2014) maintenance–retrieval model, we may expect participants to have more difficulty detecting FP than FG violations regardless of language status. The maintenance–retrieval model assumes that syntactic category information about the filler is retained during an active gap search while its lexical–semantic features may become less accessible over time. If syntactic gaps are indeed anticipated whilst semantic integration requires memory retrieval, then our participants may be more likely to erroneously accept implausible fillers than filled gaps in both of their languages.

Resource limitation approaches to L2 processing (McDonald, 2006) predict that L2 speakers may struggle more than L1 speakers when establishing long-distance dependencies and may thus be more prone to accepting ungrammatical or implausible sentences. That is, for sentences of similar structure and complexity, we would expect participants' ability to detect both FG and FP violations to be reduced by about the same extent in their L2 compared to their L1.

In contrast, an asymmetrical pattern of L2 FGD resolution difficulty is expected from the perspective of the SSH (Clahsen & Felser, 2006, 2018). If comprehenders are less likely to build fully detailed structural representations in their L2 than in their L1 during processing that include syntactic gaps, then participants' ability to detect FG violations may be reduced more dramatically in L2 vs. L1 processing than their ability to detect FP violations.

At the individual level, we expected some participants to perform better than others overall and some participants to be better at filled-gap detection than at detecting implausible fillers, and vice versa. Examining whether ungrammaticality and implausibility detection abilities were correlated within each language was motivated by the assumption that FGD resolution involves two distinct processes, syntactic and semantic filler integration. It is conceivable that these two processes are not fully independent but are both triggered automatically and operate in parallel, being equally affected by individual differences in resource limitations or other cognitive factors. In this case, we would expect to find a significant positive correlation between participants' *d'* scores in both our two German and our two English experiments. A lack of a correlation, on the other hand, would suggest that the two types of integration process can operate independently, or that for a given individual, one of the two processes is computationally easier or more automatized than the other.

If individual sensitivity to syntactic and semantic violations is shared between the L1 and L2, we should find participants' L1 and L2 accuracy scores to be positively correlated for both violation types. However, if the findings reported by Grey (2023) reflect the

existence of separate L1 and L2 processing biases or profiles, then no such correlations should be observed.

2.1. Participants

Seventy-seven native speakers of German with English as an L2 participated in our study. They were recruited through online and on-campus advertisements and received either EUR 20 or course credit for their participation. All participants had normal or corrected-to-normal vision and had no history of developmental or language disorders. They had all started learning English at school. Their English proficiency was assessed using the Quick Oxford Placement Test (Oxford University Press, 2001) and corresponded to the B2 level of the Common European Framework of Reference for Languages (range: B1–C2). Demographic information about the participants is summarized in Table 1.

Table 1. Participants' demographic information.

	Mean	**SD**	**Range**
Age	29.64	9.15	18–66 **
L2 AoA	8.91	1.81	6–14
OPT score *	47.81	6.27	33–59

* The score on the Quick Oxford Placement Test (out of 60). ** Including our oldest bilingual speaker did not alter the results of the statistical analyses.

2.2. Materials

An overview of the types of stimuli we used in our filled-gap experiments (Experiment 1a,b) and our filler plausibility experiments (Experiment 2a,b) is provided in (8) and (9).

(8) a. EXPERIMENT 1a: Filled-gap violations in German.

Michelle	*kannte*	*den*	*Lernstoff,*	*{bevor/*den}*
Michelle	knew	the	learning matter	before/*which
die	*Dozentin*	*ihrem*	*Kurs*	*diesen*
the	lecturer	her	course	this
vermittelt	*hatte.*			
taught	had			

'Michelle knew the learning matter {before/which} the lecturer taught it to her class'.

b. EXPERIMENT 1b: Filled-gap violations in English.
Michael watched the boy while/*for whom Susan was playing some music for him.

(9) a. EXPERIMENT 2a: Filler plausibility violations in German.

Kristin	*bekam*	*den*	*{Brief/#Kater},*	*den*
Kristin	received	the	letter/#tomcat	which
der	*Geliebte*	*gelesen*	*hatte.*	
the	lover	read	had	

'Kristin received {the letter/tomcat} which her lover had read'.

b. EXPERIMENT 2b: Filler plausibility violations in English.
Sharon photographed the bottle/#scientist that the kind waiter opened.
Our materials are described in more detail below.

2.2.1. Experiment 1a: Filled-Gap Violations in German

The experimental sentences for Experiment 1a were modeled after materials used by Felser and Jessen (2020). All experimental sentences were 12 words long and consisted of a main clause and a subordinate clause.

The licensing verbs in our study were ditransitive verbs that could also be used monotransitively. For instance, the verb *vermitteln* ('teach') in example (8a) has three arguments: the subject NP *die Dozentin* ('the lecturer'), the dative-marked indirect object NP *ihrem Kurs* ('her class'), and the accusative-marked direct object pronoun *diesen* ('this'), which refers back to the direct object NP *den Lernstoff* ('the learning matter') in the matrix clause. In the corresponding ungrammatical sentences, the subordinating conjunction (e.g., *bevor* 'before') was replaced by an accusative-marked relative pronoun (*den* 'which'). This pronoun signals the start of a relative clause (RC) modifying the matrix direct object NP *den Lernstoff* ('the learning matter'). The relative pronoun was always a masculine singular form, as this form is unambiguously marked for accusative case and singular number in German. Unwanted ambiguity as to the type of *wh*-extraction could thus be avoided, and the relative pronoun could only possibly be linked to a direct object gap. In our ungrammatical items, the RC's preverbal object gap position was filled by the pronoun *diesen*, however. The filled-gap violation only became obvious at the verb (*vermittelt* 'taught') as the sentence could potentially be completed in a grammatical way following *diesen* ('this'), which could also be used as a prenominal determiner introducing an adverbial expression. Note that semantically, *diesen* can be interpreted as resumptive (referring back to *den Lernstoff*), which should reduce the possibility of the filled gap being interpreted as introducing a new referent and thus causing argument or thematic role competition. That is, despite being ungrammatical, our sentences containing filled-gap violations could in principle be assigned a plausible interpretation.

All direct object NPs and all critical pronouns were masculine singular forms, since feminine and neuter pronouns are case-ambiguous in German. The experimental sentences were matched for their length. To make sure that our grammatical control sentences were indeed perceived as grammatical, we asked 14 native speakers of German who did not participate in our main experiments to rate the grammaticality of 132 sentences on a scale from 1 (=*völlig grammatisch* 'completely grammatical') to 5 (=*völlig ungrammatisch* 'completely ungrammatical'). Based on the participants' judgements, we selected the 72 best rated sentences (M = 2.1; SD = 0.25; range: 1.43–2.7) to be included in our experimental item set.

Seventy-two experimental sentence pairs were constructed and distributed across two presentation lists in a Latin square design. Additionally, 80 grammatical or ungrammatical filler sentences were created with a similar syntactic structure.

2.2.2. Experiment 1b: Filled-Gap Violations in English

The materials for the FG violation study in English were adapted from those used by Jessen et al. (2017). As in Experiment 1a, our stimulus sentences were built around optionally ditransitive verbs (example 8b). They were also of a similar length to those used in the corresponding German experiment. Our grammatical control sentences contained an adjunct clause and no gap, whilst our ungrammatical sentences contained a relative clause and a superfluous pronominal PP that could be interpreted as resumptive.

While thematic roles are indicated by morphological case in German, English lacks a proper case system. To avoid unintended ambiguity regarding the type of *wh*-extraction, in our English materials we used fronted PPs functioning as indirect objects, which could not be misinterpreted as subjects or direct objects. Note also that in contrast to German, English is an SVO language, which means that the licensing verb precedes rather than follows the filled gap here.

We pre-tested our materials with a sentence continuation questionnaire and a plausibility norming questionnaire. Thirty-four verbs admitting a *to*-PP and thirty-five verbs admitting a *for*-PP were selected from Levin (1993), around which we built a total of

205 short sentences (e.g., *Pam sold some antiques*). Thirteen native English speakers evaluated the sentences for completeness in a web-based sentence continuation questionnaire. The participants could write a continuation to the sentences if they considered it appropriate. Verbs that elicited more than 20% *to*-PP or *for*-PP completions and fewer than 70% "complete" responses were excluded. This way, we tried to prevent participants from anticipating an indirect object in our grammatical condition.

We used the remaining verbs to construct 126 sentences for a web-based grammaticality questionnaire to ensure that our control sentences were indeed deemed syntactically acceptable. Seven native speakers of English who did not further participate in this study evaluated the sentences' grammaticality on a scale ranging from 1 (="completely grammatical") to 5 (="completely ungrammatical"). Based on these judgements, we selected the 72 top-scoring sentences for our experiment (36 *for*-sentences: M: 1.57; *SD*: 0.27; range: 1.14–2; 36 *to*-sentences: M: 1.5; *SD*: 0.17; range: 1.14–1.86).

Our final set of experimental materials consisted of 72 experimental pairs as in (8) distributed across two presentation lists. Thirty-six experimental sentences involved an RC introduced by *for whom*, and 36 sentences involved an RC introduced by *to whom*. The sentences were matched for their length. Our experimental sentences were mixed with eighty (grammatical or ungrammatical) filler sentences which were structurally similar to the experimental ones. These fillers included 27 sentences that resembled our ungrammatical FG sentences in that they contained an indirect object *wh*-dependency but which were in fact grammatical, as well as several ungrammatical sentences that contained an adjunct clause. These fillers were included to prevent participants from developing any expectations about a sentence's grammaticality from the appearance of specific lexical items or clause types.

2.2.3. Experiment 2a: Filler Plausibility Violations in German

We adapted and extended experimental materials used by Felser and Jessen (2020). All experimental sentences were nine words in length and contained object RCs introduced by the accusative-marked, masculine, singular relative pronoun *den* ('which') as in example (9a) above.

All RCs modified the matrix direct object (the filler). Implausible sentences were created by manipulating the relativized direct object NP's animacy, resulting in a selectional violation that became obvious at the embedded verb. In example (9), the selectional restrictions of the verb *lesen* ('read') are violated in the case of the implausible filler *den Kater* ('the tomcat') since this verb selects for an inanimate patient argument. All experimental sentences were syntactically well-formed.

To pre-test our materials, 20 native speakers of German evaluated our sentences on a scale from 1 (=*völlig semantisch plausibel* 'completely plausible') to 5 (=*völlig semantisch unplausibel* 'completely implausible'). Only sentence pairs whose 'plausible' version scored 2.5 or less on average were included in our final materials set, provided that there was a difference of at least 2 points between a plausible sentence and its implausible counterpart.

Object nouns in the main clause were matched for their length and frequency (Baayen et al., 1993), and we counterbalanced the animacy of the object fillers across the plausible and implausible conditions. A total of 72 sentence pairs were distributed across two presentation lists and mixed with 72 filler items, which included sentences with and without a filler plausibility violation.

2.2.4. Experiment 2b: Filler Plausibility Violations in English

Our experimental sentences closely resembled the German materials described above, both in terms of their syntactic structure and their length of nine words. They consisted of a main clause followed by an RC, which was introduced by the complementizer *that*.

As in Experiment 2a, the RC modified the object of a transitive verb in the main clause, as in example (9b).

In the FP violation condition, we violated the licensing verb's (here, *opened*) selectional restrictions by manipulating the filler's animacy. All experimental sentences were grammatically correct.

In a materials evaluation pre-test, 25 native English speakers rated the plausibility of 100 sentence pairs on a scale from 1 (="completely plausible") to 5 (="completely implausible"). We applied the same exclusion criteria as in Experiment 2a, including only item pairs with a difference of two or more points between the plausible and implausible conditions. Object NPs in the main clause were matched for their length and frequency (Baayen et al., 1993) and counterbalanced for animacy. Seventy-two experimental sentence pairs were distributed across two presentation lists, set amongst 72 plausible or implausible filler items.

2.3. Procedure

All participants took part in four parallel binary judgement experiments. They were tested remotely using the online experiment platform PC Ibex (Zehr & Schwarz, 2022). The participants received detailed written instructions regarding the procedure before the presentation of each experiment. Before starting with the experiments, the participants digitally signed a consent form and completed a background questionnaire, providing demographical information and information about their health (eyesight, history of language impairments and neurological diseases), language use, and linguistic background. The experiments were presented in a pseudo-randomized order to control for potential bias and achieve balanced exposure. Participants had up to one week to complete the full set of experiments and were asked to complete the experiments in the prescribed order. Each experimental list contained 36 correct (plausible/grammatical) and 36 incorrect (implausible/ungrammatical) sentences, plus 80 (Experiment 1a,b) or 72 (Experiment 2a,b) filler sentences, making a total of 152 (Experiment 1a,b) or 144 (Experiment 2a,b) items. The stimuli in each list were counterbalanced using a Latin square design and randomized for each participant.

The stimuli were presented in black font on a white background in Times New Roman (12 pts) on the computer screen. We used a moving-window presentation mode to display the sentences, with one word appearing at a time and the rest of the sentence represented by dashes. Each word was displayed for 450 ms, after which it disappeared, and 100 ms later the next word appeared on the screen. At the end of a sentence, the dashes disappeared, and participants were shown a response prompt. For grammaticality judgements, the prompt was *Ist dieser Satz grammatisch korrekt?/Is this sentence grammatical?* and for plausibility judgements, it was *Ist dieser Satz semantisch plausibel?/Is this sentence semantically plausible?* The question was visible for 3000 ms and disappeared when no response was given (timeout). Participants provided their answers by pressing either the designated "yes" ("J") or "no" key ("F") on their keyboard. Participants' judgements and response times (RTs) were recorded.

Participants were instructed to read the stimulus sentences carefully and to respond as quickly and as accurately as possible. Prior to the experiments, participants read an explanation of the terms "grammatical" and "plausible" to ensure consistent interpretation. At the beginning of each experiment, participants were presented with five practice items. The participants could take a break after every trial whenever they wished. Each experiment lasted approximately 30 min.

2.4. Data Analysis

We fitted four mixed-effects logistic regression models (binomial family) with the lme4 package in R (Bates et al., 2015). Note that we will only analyze and discuss the response data and not the RT data as we consider participants' judgements more informative for answering our research questions. Raw RTs are presented for completeness only. Responses were coded as 0 and 1 in the variable *Answer*, which stood for the reply "grammatical"/"plausible" and the reply "ungrammatical"/"implausible", respectively. Expected responses were also coded as 0 and 1 in the factor *Expected Response*, following the same logic as the coding of the variable *Answer*. The factor *Manipulation Type* was sum-coded as −0.5 for the FP experiments and 0.5 for the FG experiments. The factor *Language* was contrast-coded as −0.5 for German and 0.5 for English. Continuous predictors were mean-centered (i.e., OPT score). Random slopes for *Language* and *Expected Response* were added to the model.

Signal Detection Theory (Pastore & Scheirer, 1974) provides the necessary tools for measuring participants' ability to differentiate between "signal" trials (here, ungrammatical or implausible sentences) and "noise" trials (here, grammatical or plausible sentences). For our correlation analyses, we calculated individual *d'* scores as an index of sensitivity to a given experimental manipulation using the psycho package in R (Makowski, 2018). This measure allowed us to assess comprehenders' sensitivity to our FG and FP manipulations whilst eliminating potential *yes* or *no* response biases in individual participant data. In order to compute *d'* scores, we calculated the rates of hits, false alarms, misses, and correct rejections. Rejecting a sentence with an FG or FP violation would be a "hit" and accepting a grammatical or a plausible sentence would be a "correct rejection". Erroneously accepting a sentence with a violation would be a "miss", while rejecting a violation-free sentence would constitute a "false alarm". *D'* scores of 0 indicate no sensitivity, and the higher a participant's *d'* score, the higher their sensitivity.

3. Results

Table 2 summarizes the raw accuracy and RT data (for correct responses). The overall accuracy rates were rather high, confirming that our participant group performed the task attentively and was generally sensitive to both types of violation in both languages.

Table 2. Mean accuracy scores (in percent) and RTs (in milliseconds, with SDs in parentheses) for the FG and FP speeded judgement experiments in German and English split by language and condition.

	German		**English**	
	No FG *	**FG ****	**No FG**	**FG**
Mean accuracy	92.73	95.34	75.39	77.44
Range	58.33–100	61.11–100	41.67–100	41.67–100
Mean RT (sd)	489 (435)	434 (403)	569 (492)	543 (484)
	Plausible	**Implausible**	**Plausible**	**Implausible**
Mean accuracy	89.19	93.68	83.33	83.76
Range	61.11–100	63.89–100	44.4–100	50–100
Mean RT (sd)	548 (478)	456 (405)	683 (526)	628 (529)

* Sentences without a filled-gap violation (grammatical). ** Sentences containing a filled-gap violation (ungrammatical).

As might be expected, our participant group showed higher accuracy scores in their L1 than in their L2 in both the FG and FP experiments. While participants' response accuracy was slightly higher in the FG than in the FP experiment in German, the opposite numerical pattern was seen in English.

3.1. Between-Language Analyses of the Response Data

3.1.1. Filled-Gap Experiments

We fitted a probit mixed model to estimate the effects of Expected Response and Language on the answers of our participants in the FG experiment. The summary of the model's output is presented in Table 3 (log odds coefficients). We observed a significant effect of Expected Response, suggesting that our participants reliably differentiated between grammatical and ungrammatical sentences in both languages. The significant main effect of Language reflects lower accuracy in the L2 than in the L1 experiment. The significant negative interaction of Expected Response and Language indicates that the effect of Expected Response affected the responses differently depending on the language of presentation. Following up on this interaction, we found that participants were more likely to incorrectly reject grammatical sentences in L2 English compared to their overall likelihood of making this error across languages (Expected Response 0 × Language, $b = 0.7601$, $SE = 0.1041$, $z = 7.30$, $p < .001$, and Expected Response 1 × Language, $b = -1.1128$, $SE = 0.1424$, $z = -7.82$, $p < .001$).

Table 3. Model output for the fixed factors in response data from the FG experiments.

Fixed Effects	*b*	*SE*	*z*	*p*
Intercept	−1.2044	0.0681	−17.68	<.001
Language	0.7601	0.1041	7.30	<.001
Expected Response	2.6774	0.1145	23.39	<.001
Expected Response × Language	−1.8728	0.2082	−9.00	<.001

Formula in R: answer ~1 + exp.resp * language + (1 + language/exp.resp | subject).

3.1.2. Filler Plausibility Experiments

The same model was fitted for the data from the FP experiments (see Table 4 for a summary). We again observed a significant effect of Expected Response, confirming that participants differentiated between plausible and implausible stimuli. The main effect of Language indicates that participants' ability to detect implausible fillers was significantly reduced in the L2 compared to the L1. The significant interaction of Expected Response and Language suggests a difference in L1/L2 response patterns depending on the presence or absence of a violation. Further analysis of this interaction revealed that discrimination between implausible and plausible items was different in the two languages. Completing the task in the L2 increased the likelihood of incorrectly discarding plausible sentences (i.e., responding with "implausible" to plausible items) (Expected Response 0 × Language, $b = 0.2873$, $SE = 0.0655$, $z = 4.39$, $p < .001$, and Expected Response 1 × Language, $b = -0.6977$, $SE = 0.0938$, $z = -7.44$, $p < .001$).

Table 4. Model output for the fixed factors in response data from the FP experiments.

Fixed Effects	*b*	*SE*	*z*	*p*
Intercept	−1.1687	0.0471	−24.81	<.001
Language	0.2873	0.0646	4.39	<.001
Expected Response	2.5659	0.0889	28.85	<.001
Expected Response × Language	−0.9850	0.1175	−8.39	<.001

Formula in R: answer ~1 + exp.resp * language + (1 + language/exp.resp | subject).

3.2. Within-Language Analyses of the Response Data

3.2.1. German Experiments

To estimate the effect of Expected Response and Manipulation Type on the behavior of our participants in their L1, we fitted a probit mixed model. A summary of the output is presented in Table 5. The effect of Expected Response turned out to be significant,

indicating that participants reliably differentiated between grammatical/ungrammatical and plausible/implausible items in their native language. We additionally observed a significant main effect of Manipulation Type, reflecting higher accuracy in the FG than in the FP experiment. The interaction of Expected Response and Manipulation Type was also significant. Following up on this interaction revealed that the likelihood of correctly rejecting sentences containing filled gaps was higher than the likelihood of correct rejections across both experiments together (Expected Response 0 × Manipulation Type, $b = -0.2390$, $SE = 0.0769$, $z = -3.11$, $p = .002$, and Expected Response 1 × Manipulation Type, $b = 0.2795$, $SE = 0.1279$, $z = 2.19$, $p = .029$).

Table 5. Model output for the fixed factors in response data from the L1 experiments.

Fixed Effects	*b*	*SE*	*z*	*p*
Intercept	−1.4378	0.0556	−25.85	<.001
Manipulation Type	−0.2390	0.0769	−3.11	.002
Expected Response	3.3109	0.1028	32.22	<.001
Expected Response × Manipulation Type	0.5186	0.1244	4.17	<.001

Formula in R: answer ~1 + exp.resp * m.type + (1 + m.type/ exp.resp | subject).

3.2.2. English Experiments

The output of the probit mixed model for the English data is presented in Table 6. The significant main effect of Expected Response indicates that our participants also successfully differentiated between grammatical/plausible and ungrammatical/implausible items in their L2. The main effect of Manipulation Type reflects lower accuracy in the FG than in the FP experiment. The interaction of Expected Response and Manipulation Type was only marginally significant.

Table 6. Model output for the fixed factors in response data from the L2 experiments.

Fixed Effects	*b*	*SE*	*z*	*p*
Intercept	−0.9236	0.0615	−15.03	.001
Manipulation Type	0.2009	0.0886	2.27	.023
Expected Response	1.9028	0.1076	17.68	<.001
Expected Response × Manipulation Type	−0.3355	0.1825	−1.84	.066

Formula in R: answer ~1 + exp.resp * m.type + (1 + m.type/exp.resp | subject).

3.3. Correlation Analyses

We conducted a series of correlation analyses of participants' *d'* scores to identify individual sensitivity to the two types of FGD violation. The results of these analyses are visualized in Figure 1. Analyses at the between-language level yielded a significant correlation between participants' *d'* scores for FP violations in English and German ($R = .43$, $p < .001$) (Experiments 2a and 2b, Figure 1a). In contrast, participants' *d'* scores for FG violations did not correlate between the two languages ($R = .15$, $p = .19$) (Experiments 1a and 1b, Figure 1b). A significant correlation was obtained between participants' *d'* scores for FG and FP violations in German ($R = .53$, $p < .001$) (Experiments 1a and 2a, Figure 1c). However, no such correlation was observed for the English data ($R = .22$, $p = .064$) (Experiments 1b and 2b, Figure 1d).

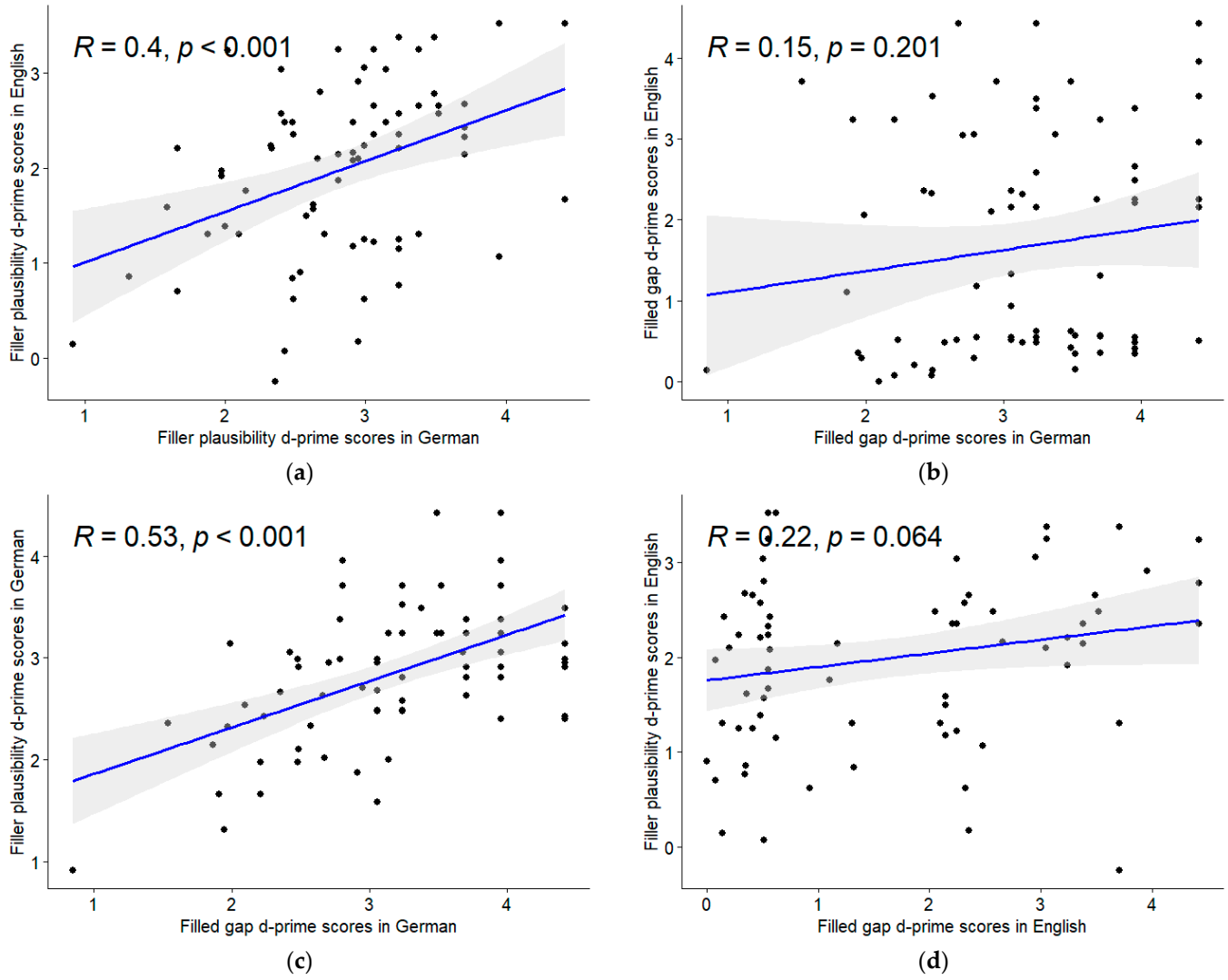

Figure 1. Correlation plots. (**a**) Correlation between *d*′ scores for the filler plausibility experiments in L1 and L2; (**b**) correlation between *d*′ scores for the filled gap experiments in L1 and L2; (**c**) correlation between *d*′ scores for the FP and FG experiments in L1; (**d**) correlation between *d*′ scores for the FP and FG experiments in L2. The firm lines are regression lines. Individual points stand for individual participants. The grey band indicates a 95% CI.

3.4. Exploratory Analyses

As our experimental sentences were rather complex, we carried out additional analyses to examine whether participants' performance in the L2 was influenced by their English proficiency. We correlated participants' *d*′ scores in Experiments 1b and 2b with their OPT scores. The results of these correlation analyses are presented in Figure 2. While participants' OPT scores were positively correlated with their sensitivity to FP violations ($R = .64$, $p < .001$; Figure 2a), their sensitivity to FG violations did not correlate with their proficiency scores ($R = .11$, $p = .348$; Figure 2b). Thus, while participants with a higher L2 proficiency were better at detecting implausible fillers than participants with a lower L2 proficiency, a higher L2 proficiency did not lead to an improved ability to detect filled gaps.

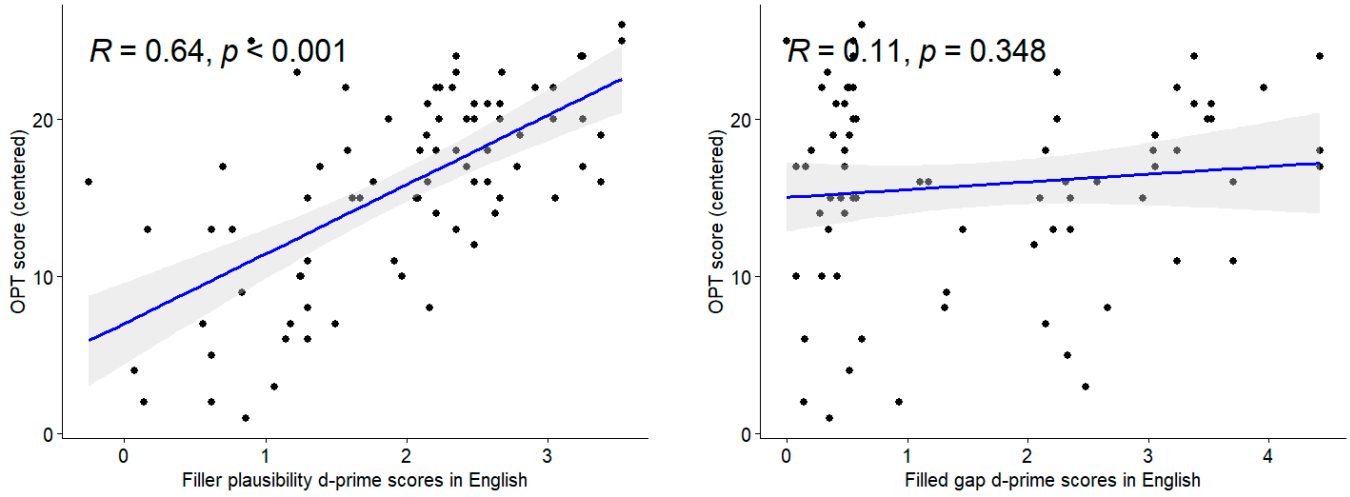

Figure 2. Correlation plots. (**a**) Correlation between individual OPT scores (centered) and *d'* scores in the L2 filler plausibility experiment. (**b**) Correlation between individual OPT scores (centered) and *d'* scores in the L2 filled-gap experiment. The firm lines are regression lines. Individual points stand for individual participants. The grey band indicates a 95% CI.

4. Discussion

We asked how language status would affect bilingual speakers' sensitivity to filled-gap and filler plausibility violations in L1 German and L2 English and whether individual sensitivity scores for different types of FGD violation would be correlated within and/or across the two languages. We assumed differences in the ability to detect filled-gap vs. filler plausibility violations to be indicative of different processing biases. While the sensitivity to filled gaps likely reflects syntactic integration processes, the sensitivity to semantically inappropriate fillers reflects semantic integration ability. Stronger sensitivity to one or the other might then be taken to indicate a bias towards either gap-driven or semantically driven FGD resolution.

Unlike previous L1/L2 FGD resolution studies, which investigated either syntactic gap effects or filler mismatch effects across different L1 and L2 speaker groups, we combined two experimental paradigms and tested the same individuals in both their L1 and L2. Our results show that at the group level, our participants performed better in their L1 than in their L2 for both violation types. We also found participants' performance patterns to be asymmetrical, with opposite accuracy patterns across violation types in German and English. While participants showed significantly higher accuracy in the filled-gap than in the filler plausibility experiment in German, their accuracy was significantly lower in the FG than in the FP experiment in English.

As was expected, we found considerable individual variability in participants' sensitivity to different violation types (Figure 1). We also observed interesting asymmetries in our individual-level analyses. The participants' ability to detect FP violations—but not their ability to detect FG violations—was positively correlated in their L1 and L2, and their sensitivity scores for the two types of FGD violations were correlated only in their L1 but not in their L2. In the following subsections, we will discuss the implications of these findings in turn.

4.1. Group-Level Effects of Language Status

Our finding that our participant group was less accurate in their L2 than in their L1 is unsurprising and in line with the assumption that L2 processing is more cognitively demanding and less automatized than L1 processing (McDonald, 2006). This assumption

cannot by itself account for the asymmetrical performance patterns that we observed across our four experiments, however.

Recall that in German, we found participants' judgement accuracy to be higher in the FG experiment (Experiment 1a) than in the FP experiment (Experiment 2a). This result is in line with the maintenance–retrieval model (Wagers & Phillips, 2014), which postulates that semantic and syntactic cues differ in their longevity during FGD formation. Participants' ability to anticipate the gap may have made filled-gap detection fairly easy for them. Anticipating the licensing verb, on the other hand, was impossible, and retrieving the filler's semantic features upon encountering the verb may have been more difficult in comparison to recognizing filled syntactic gaps.

The opposite accuracy pattern was seen for the corresponding English experiments, however. Here, participants' responses in the FG experiment (Experiment 1b) were considerably less accurate than their responses in the FP experiment (Experiment 2b). This is unexpected from the perspective of the maintenance–retrieval model but in line with the prediction made by the SSH. The SSH assumes that L2 comprehenders are less syntactically driven than L1 comprehenders but can compensate for this by utilizing semantic and other non-structural cues to interpretation. For FGD resolution, this hypothesis predicts that L2 comprehenders should detect implausible fillers more easily and more reliably than filled gaps. This is precisely what we found. Our results match the L1/L2 processing differences that were observed in Felser et al.'s (2012) eye-movement monitoring study and also fit with earlier findings showing that structural gap effects, but not filler plausibility effects, are often absent or non-nativelike in L2 comprehension (see Section 1.2).

Although we used the same sentence types and the same type of verb in both FG experiments, one difference between our German (Experiment 1a) and English (Experiment 1b) FG materials was that the filled gap appeared preverbally in German but postverbally in English. Participants were thus able to evaluate the semantic fit between the filler and verb prior to encountering the filled gap in English (Experiment 1b) but not in German (Experiment 1a). If we follow Nicol (1993) and others in assuming that FGD resolution involves both semantic (i.e., verb-based) and syntactic (i.e., gap-based) integration processes, then the fact that successful semantic filler integration could already be carried out at the verb might have increased the likelihood of participants erroneously accepting English filled-gap violations as grammatical. Recall that the filler was always semantically compatible with the licensing verb in our ungrammatical FG sentences. Together with the fact that the constituent filling the gap could be interpreted resumptively, this means that our ungrammatical FG sentences were nevertheless interpretable and that the appearance of the filled gap would not necessarily disrupt successful incremental semantic interpretation.

Recall that in our German filled-gap materials, the point at which the ungrammatically was confirmed coincided with the point at which the filler could be fully integrated semantically (i.e., at the licensing verb). As our study was not designed to measure position-specific or incremental word-by-word processing, we cannot determine whether the SOV/SVO typological difference between the two languages had any influence on the likelihood of our participants detecting filled gaps.

In short, our finding that participants had comparatively less difficulty identifying implausible fillers than filled gaps in English fits with the assumption that FGD resolution is more semantically than syntactically driven in L2 comprehension (Felser et al., 2012; Marinis et al., 2005).

4.2. Individual Differences

By glossing over individual differences in L1 and L2 performance, traditional L1 vs. L2 group-level analyses may yield questionable "native" vs. "nonnative" process-

ing dichotomies (Grey, 2023). To examine both inter- and intra-individual variability in participants' sensitivity to FGD violations and how individual participants' performance within and between their two languages might be related, we carried out a series of correlation analyses on their *d'* scores (Figure 1). Regarding the within-language performance, our correlation analyses demonstrated that participants' sensitivity to the two types of violation was correlated in their L1. That is, participants who were good at detecting FG violations in German were also good at detecting FP violations in German. It is conceivable that the observed within-language variability in participants' general ability to identify FGD violations in German was influenced by cognitive factors such as WM or attentional control that were not of interest to the current study. However, no positive or negative correlation between FG and FP sensitivity was observed for English. Language status thus constrains intra-individual variability such that participants' performance across our FG and FP experiments was less variable in their L1 than in their L2. The lack of a correlation between participants' *d'* scores in our two English experiments provides evidence that syntactic and semantic integration processes are distinct and not necessarily linked in the same individual.

Given that the same individuals were tested in both German and English, any individual cognitive (or other not language-related) differences should affect FGD resolution in both languages equally. We did indeed find a positive correlation for participants' ability to detect FP violations in L1 German and L2 English. This suggests that the individual sensitivity to semantic mismatches in FGD resolution carries over from the L1 to the L2. We found no correlation for participants' ability to detect FG violations in their two languages, however. That is, individual sensitivity to filled gaps in the L1 does not predict sensitivity to filled gaps in the L2.

The observed correlation asymmetries indicate that, while bilinguals have similar semantic integration abilities in both L1 and L2 comprehension, their syntactic integration ability varies independently in their L1 and L2. Regarding the question of whether processing abilities or biases are shared across an individual's languages, our results suggest that semantic but not syntactic processing ability is shared. To explore whether the participants' ability to detect filled gaps and/or implausible fillers in the L2 was linked to their English proficiency as measured by the Quick OPT, we correlated their *d'* scores from Experiments 1b and 2b with their proficiency scores. While participants' ability to identify implausible fillers increased with increasing proficiency (also compare Dallas et al., 2013), a higher English proficiency did not improve participants' sensitivity to filled gaps in their L2.

The selective L1/L2 correlation we found for semantic violations was not seen in Grey's (2023) brain response data. Unlike what we observed, Grey found that neither participants' brain responses to grammatical violations nor their brain responses to semantic violations were correlated between L1 English and L2 Spanish. Clearly, further studies are needed on different language combinations and different linguistic phenomena to explore for which linguistic domains and under what conditions bilingual speakers' processing ability or biases are shared between their languages.

Considering the observed within- and between-language variability and the fact that some individuals performed less well in their L1 than in their L2, we may ask whether setting up a simple dichotomy between "native" and "nonnative" processing is justified and, indeed, helpful. Individuals may differ in the extent to which comprehension is semantically or syntactically driven in their L1 and their L2. Theoretical approaches to L2 processing which assume that L1/L2 differences are gradient in that they allow for different types of cues or constraints to be weighted differently in the L1 and L2 (Clahsen & Felser, 2018; Cunnings, 2017; MacWhinney, 2008) may need to be extended so as to allow for both inter- and intra-individual processing differences to be captured. Determining individual

processing profiles across different languages and phenomena should allow us to capture individual variability in bilingual processing and identify the factors that constrain it.

5. Conclusions

In four parallel speeded judgement experiments, we found bilinguals' ability to detect FGD violations to be reduced in L2 relative to L1 comprehension, but with their ability to detect filled syntactic gaps disproportionally reduced in the L2 in comparison to their ability to detect implausible fillers. This pattern of results confirms the prediction of the SSH, which claims that L2 comprehension is less likely than L1 comprehension to be based on the computation of detailed grammatical representations. In terms of theoretical modeling, this putative L1/L2 difference can be accounted for by assuming that structural and semantic information is differently weighted in L1 and L2 comprehension. We also observed considerable individual variability in participants' response patterns across their two languages, however. This argues for a more nuanced approach to determining L1/L2 processing differences than is usually adopted and calls into question the traditional distinction between "native" and "nonnative" linguistic performance patterns.

Author Contributions: Conceptualization, C.F.; methodology, A.T. and C.F.; software, A.T.; validation, A.T.; formal analysis, A.T.; investigation, A.T.; resources, A.T. and C.F.; data curation, A.T.; writing—original draft preparation, A.T. and C.F.; writing—review and editing, A.T. and C.F.; visualization, A.T.; supervision, C.F.; project administration, C.F.; funding acquisition, C.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–Project ID 317633480–SFB 1287.

Institutional Review Board Statement: This study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the University of Potsdam (Reference number 37/2011).

Informed Consent Statement: Informed consent was obtained from all participants in the study.

Data Availability Statement: The data are available in the study's OSF repository at https://osf.io/rk6sy/ (accessed on 11 January 2025).

Acknowledgments: We thank Friederike Schubert and Marian Jenke for assisting with the materials creation. Special thanks go to Anna Laurinavichyute and Carlotta Zona for their statistical advice and to Maximilian Rabe for sharing his expertise on Signal Detection Theory. We extend our gratitude to all our participants.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Aldwayan, S., Fiorentino, R., & Gabriele, A. (2010). Evidence of syntactic constraints in the processing of wh-movement. *Research in Second Language Processing and Parsing*, *53*, 65–86.

Al-Maani, A., Sloggett, S., Grillo, N., & Marsden, H. (2024). Testing for proficiency effects and crosslinguistic influence in L2 processing: Filler-gap dependencies in L2 English by Jordanian Arabic and Mandarin speakers. *Studies in Second Language Acquisition*, *46*(2), 564–580. [CrossRef]

Baayen, R. H., Piepenbrock, R., & Van Rijn, H. (1993). *The CELEX lexical database (CD-ROM) Philadelphia*. Linguistic Data Consortium, University of Pennsylvania.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv*, arXiv:1506.04967.

Berghoff, R. (2022). L2 Processing of filler-gap dependencies: Attenuated effects of naturalistic L2 exposure in a multilingual setting. *Second Language Research*, *38*(2), 373–393. [CrossRef]

Berghoff, R. (2023). Wh-dependency processing in a naturalistic exposure context: Sensitivity to abstract syntactic structure in high-working-memory L2 Speakers. *Studies in Second Language Acquisition*, *45*(2), 586–98. [CrossRef]

Blackwell, A., Bates, E., & Fisher, D. (1996). The time course of grammaticality judgement. *Language and Cognitive Processes*, *11*(4), 337–406. [CrossRef]

Boxell, O., & Felser, C. (2017). Sensitivity to parasitic gaps inside subject islands in native and non-native sentence processing. *Bilingualism: Language and Cognition*, *20*(3), 494–511. [CrossRef]

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics*, *27*(1), 3–42. [CrossRef]

Clahsen, H., & Felser, C. (2018). Some notes on the shallow structure hypothesis. *Studies in Second Language Acquisition*, *40*(3), 693–706. [CrossRef]

Covey, L., Fiorentino, R., & Gabriele, A. (2024). Island sensitivity in L2 learners: Evidence from acceptability judgments and event-related potentials. *Second Language Research*, *40*(1), 19–50. [CrossRef]

Cunnings, I. (2017). Parsing and working memory in bilingual sentence processing. *Bilingualism: Language and Cognition*, *20*(4), 659–78. [CrossRef]

Dallas, A., DeDe, G., & Nicol, J. (2013). An event-related potential (ERP) investigation of filler-gap processing in native and second language speakers. *Language Learning*, *63*(4), 766–99. [CrossRef]

Dong, Z. R., Han, C., Hestvik, A., & Hermon, G. (2023). L2 Processing of filled gaps: Non-native brain activity not modulated by proficiency and working memory. *Linguistic Approaches to Bilingualism*, *13*(6), 767–800. [CrossRef]

Felser, C., Cunnings, I., Batterham, C., & Clahsen, H. (2012). The timing of island effects in nonnative sentence processing. *Studies in Second Language Acquisition*, *34*(1), 67–98. [CrossRef]

Felser, C., & Jessen, A. (2020). Brain responses elicited by implausible fillers and filled object gaps in German. *Typical and Impaired Processing in Morphosyntax*, *64*, 75.

Felser, C., & Roberts, L. (2007). Processing wh-dependencies in a second language: A cross-modal priming study. *Second Language Research*, *23*(1), 9–36. [CrossRef]

Frazier, L. (1987). Theories of sentence processing. In J. L. Garfield (Ed.), *Modularity in knowledge representation and natural-language understanding* (pp. 291–308). The MIT Press.

Garnsey, S. M., Tanenhaus, M. K., & Chapman, R. M. (1989). Evoked potentials and the study of sentence comprehension. *Journal of Psycholinguistic Research*, *18*(1), 51–60. [CrossRef] [PubMed]

Grey, S. (2023). Variability in native and nonnative language: An ERP study of semantic and grammar processing. *Studies in Second Language Acquisition*, *45*(1), 137–66. [CrossRef]

Hartsuiker, R. J., Pickering, M. J., & Veltkamp, E. (2004). Is syntax separate or shared between languages? Cross-linguistic syntactic priming in Spanish-English bilinguals. *Psychological Science*, *15*(6), 409–14. [CrossRef]

Hestvik, A., Maxfield, N., Schwartz, R. G., & Shafer, V. (2007). Brain responses to filled gaps. *Brain and Language*, *100*(3), 301–16. [CrossRef] [PubMed]

Hopp, H. (2015). Individual differences in the second language processing of object–subject ambiguities. *Applied Psycholinguistics*, *36*(2), 129–73. [CrossRef]

Jessen, A., & Felser, C. (2019). Reanalysing object gaps during non-native sentence processing: Evidence from ERPs. *Second Language Research*, *35*(2), 285–300. [CrossRef]

Jessen, A., Festman, J., Boxell, O., & Felser, C. (2017). Native and non-native speakers' brain responses to filled indirect object gaps. *Journal of Psycholinguistic Research*, *46*, 1319–1338. [CrossRef] [PubMed]

Johnson, A., Fiorentino, R., & Gabriele, A. (2016). Syntactic constraints and individual differences in native and non-native processing of wh-movement. *Frontiers in Psychology*, *7*, 549. [CrossRef] [PubMed]

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual differences in language acquisition and processing. *Trends in Cognitive Sciences*, *22*(2), 154–69. [CrossRef]

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. University of Chicago Press.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*(3), 375–419. [CrossRef] [PubMed]

MacWhinney, B. (2008). A unified model. In *Handbook of cognitive linguistics and second language acquisition* (pp. 351–381). Routledge.

Makowski, D. (2018). The Psycho Package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, *3*(22), 470. [CrossRef]

Marinis, T., Roberts, L., Felser, C., & Clahsen, H. (2005). Gaps in second language sentence processing. *Studies in Second Language Acquisition*, *27*(1), 53–78. [CrossRef]

McDonald, J. L. (2006). Beyond the critical period: Processing-based explanations for poor grammaticality judgment performance by late second language learners. *Journal of Memory and Language*, *55*(3), 381–401. [CrossRef]

Miller, A. K. (2015). Intermediate traces and intermediate learners: Evidence for the use of intermediate structure during sentence processing in second language French. *Studies in Second Language Acquisition*, *37*(3), 487–516. [CrossRef]

Nakano, Y., Felser, C., & Clahsen, H. (2002). Antecedent priming at trace positions in Japanese long-distance scrambling. *Journal of Psycholinguistic Research*, *31*, 531–571. [CrossRef]

Nicol, J. L. (1993). Reconsidering Reactivation. In G. Altmann, & R. Shillcock (Eds.), *Cognitive models of speech processing* (pp. 321–350). Taylor and Francis.

Omaki, A., & Schulz, B. (2011). Filler-gap dependencies and island constraints in second-language sentence processing. *Studies in Second Language Acquisition*, *33*(4), 563–88. [CrossRef]

Oxford University Press. (2001). *Quick placement test*. Oxford University Press.

Pastore, R. E., & Scheirer, C. J. (1974). Signal detection theory: Considerations for general application. *Psychological Bulletin*, *81*(12), 945. [CrossRef]

Pickering, M., & Barry, G. (1991). Sentence processing without empty categories. *Language and Cognitive Processes*, *6*(3), 229–59. [CrossRef]

Pliatsikas, C., & Marinis, T. (2013). Processing empty categories in a second language: When naturalistic exposure fills the (intermediate) gap. *Bilingualism: Language and Cognition*, *16*(1), 167–82. [CrossRef]

Smolensky, P., Goldrick, M., & Mathis, D. (2014). Optimization and quantization in gradient symbol systems: A framework for integrating the continuous and the discrete in cognition. *Cognitive Science*, *38*(6), 1102–38. [CrossRef] [PubMed]

Stowe, L. A. (1986). Parsing WH-constructions: Evidence for on-line gap location. *Language and Cognitive Processes*, *1*(3), 227–45. [CrossRef]

Tanner, D., McLaughlin, J., Herschensohn, J., & Osterhout, L. (2013). Individual differences reveal stages of L2 grammatical acquisition: ERP evidence. *Bilingualism: Language and Cognition*, *16*(2), 367–82. [CrossRef]

Wagers, M. W., & Phillips, C. (2014). Going the distance: Memory and control processes in active dependency construction. *The Quarterly Journal of Experimental Psychology*, *67*(7), 1274–1304. [CrossRef]

Williams, J. N. (2006). Incremental interpretation in second language sentence processing. *Bilingualism: Language and Cognition*, *9*(1), 71–88. [CrossRef]

Williams, J. N., Möbius, P., & Kim, C. (2001). Native and non-native processing of English wh-questions: Parsing strategies and plausibility constraints. *Applied Psycholinguistics*, *22*(4), 509–540. [CrossRef]

Yadav, H., Paape, D., Smith, G., Dillon, B. W., & Vasishth, S. (2022). Individual differences in cue weighting in sentence comprehension: An evaluation using approximate Bayesian computation. *Open Mind*, *6*, 1–24. [CrossRef] [PubMed]

Zehr, J., & Schwarz, F. (2022). *PennController for internet based experiments (IBEX)*. Center for Open Science.