L2 Transfer of L1 Island-Insensitivity:
The case of Norwegian

Dave Kush
Department of Linguistics, University of Toronto, Canada
Centre for French & Linguistics, University of Toronto Scarborough, Canada
NTNU – Norwegian University of Science and Technology

Anne J. Dahl
NTNU – Norwegian University of Science and Technology

Corresponding Author:
Dave Kush
Institutt for språk og litteratur
NTNU Dragvoll
7491 Trondheim
dave.kush@ntnu.no

Abstract

Norwegian allows filler-gap dependencies into embedded questions, which are islands for filler-gap dependency formation in English. We ask whether there is evidence that Norwegian learners of English transfer the functional structure that permits island violations from their L1 to their L2. In two acceptability judgment studies, we find that Norwegians are more likely to accept 'island-violating' filler-gap dependencies in L2 English if the corresponding filler-gap dependency is acceptable in Norwegian: Norwegian learners variably accept English sentences with dependencies into embedded questions, but not into subject phrases. These results are consistent with models that permit transfer of abstract functional structure (e.g. Eubank 1993/1994; Schwartz & Sprouse 1996; Westergaard 2019). Norwegians are still less likely to accept filler-gap dependencies into English embedded questions than Norwegian embedded questions. We interpret the latter finding as evidence that despite transfer, Norwegian speakers may partially restructure their L2 English analysis. We discuss how indirect positive evidence may play a role in helping learners restructure.

Keywords: Syntactic Transfer, island effects, Norwegian, filler-gap dependencies, indirect evidence

This paper addresses L1 transfer in the acquisition of filler-gap dependencies in adult second language (L2) acquisition. We ask whether Norwegian learners of English transfer acceptable filler-gap dependencies from their L1 Norwegian to their L2 English, including dependencies that are unacceptable (and therefore unattested) in English. We also consider whether and how Norwegians might learn that English is more restrictive than Norwegian.

Norwegian and English allow long-distance filler-gap dependencies into embedded declarative clauses. For example, the relative clause (RC) head *the signals/signalene* can be interpreted as either the direct object (1a, 2a) or subject (1b, 2b) of an embedded verb.

(1) a. Those were the signals$_i$ that the sailors said [(that) folks could understand ___$_i$].
b. Those were the signals$_i$ that the sailors said [___$_i$ meant danger].

(2) a. Det var signal-ene som sjømenn-ene sa [(at) folk kunne forstå ___$_i$].
That was signal-DEF.PL that seamen-DEF.PL said that folks could understand
`Those were the signals that the sailors said that folks could understand.'
b. Det var signal-ene som sjømenn-ene sa [(at) __$_i$ betydde fare].
That was signal-DEF.PL that seamen-DEF.PL said that meant danger
`Those were the signals that the seamen said meant danger.'

Norwegian and English differ, however, in subtle ways. Embedded questions are *islands* in English in that they block filler-gap dependency formation (Chomsky, 1977; Sprouse, Wagers & Phillips, 2012). Attempting to associate the filler *the signals* with the embedded verbs in (3) results in unacceptability. In Norwegian embedded questions are not islands (Maling & Zaenen, 1982). It is acceptable to associate the filler *signalene* with the embedded gaps in (4).

(3) a. *Those were the signals$_i$ that the sailors knew [**who** could understand __ $_i$].
b. *Those were the signals$_i$ that the sailors knew [**what** __$_i$ meant].

(4) a. Det var signal-ene$_i$ som sjømennene visste [**hvem** som kunne forstå ___$_i$].
That was signal-DEF.PL that seamen-DEF.PL knew who EXPL could understand
`Those were the signals that the sailors knew who could understand.'
b. Det var signal-ene$_i$ som sjømennene visste [**hva** __$_i$ betydde].
That was signal-DEF.PL that seamen-DEF.PL knew what meant
`Those were the signals that the sailors knew what meant.'
~ `Those were the signals that the sailors knew the meaning of.'

In the present study, we investigate if the acceptability of sentences like (4b) leads native Norwegian speakers to accept sentences like (3b) in their L2 English.

We expect Norwegians to accept sentences like (3b) if they inappropriately *transfer* those features of their L1 grammar that render embedded questions non-islands to L2 English. What could such features be? Under many generative syntactic analyses 'long-distance' movement out of an embedded clause as in (1) and (2) requires successive-cyclic movement through the left-periphery of the embedded clause (e.g., Chomsky 1977, 2000). In languages like English, this movement uses the specifier of the complementizer phrase (henceforth spec,CP) as an intermediate landing site. Ordinary declarative clauses are not islands, because spec,CP is empty, allowing the moved element to transit through. Embedded questions are islands because spec,CP is already occupied by the wh-phrase (*who/what* in 3a,b), so an intermediate stop-over is blocked.

Cross-linguistic differences in the islandhood of embedded questions are assumed to reflect parametric variation in the functional structure of the left-periphery of the clause (e.g., Rizzi 1982;

Reinhart 1983).[1] For the sake of concreteness we make use of a specific proposal made for Mainland Scandinavian languages like Norwegian: Recent work argues that such languages have multiple specifiers in the complementizer domain that would permit successive-cyclic movement through the edge of an embedded question (e.g., Vikner et al. 2017; Lindahl 2017, 2019; Kush, Lohndal & Sprouse 2018, 2019). The relevant specifiers are generated by an extra functional head (e.g. the head *c* under Vikner and colleagues' proposal).

Under this analysis, Norwegians would treat English embedded questions as non-islands if they transfer the extra functional structure of their L1 complementizer domain to English. As we discuss below, whether such transfer is possible is a point of disagreement between models of L2 acquisition. In our investigation, we address three inter-related theoretical questions at the intersection of second-language influence and learnability:

1. To what extent does L1 functional structure transfer to L2?
2. Are L1 features transferred to L2 in a conservative fashion?
3. How do L2 learners *restructure* after erroneous L1 transfer?

We discuss each question in turn.

What Can Transfer?

Cases of L1-L2 transfer are well documented. For example, learners often produce or accept L1 word order patterns that are ungrammatical in L2 (Trahey & White, 1993; White, 1991; Ayoun, 1999; Westergaard, 2003; Rankin, 2012). Such instances suggest that L2 learners use some aspects of their L1 (as a starting point) to analyze their L2, but exactly what transfers is a matter of considerable debate.

Models of L2/Ln acquisition disagree on the degree to which L1 functional structure transfers (see Rothman, González Alonso & Puig Mayenco, 2019 for review). The *Minimal Trees* approach of Vainikka & Young-Scholten (1994, 1996, 2006) admits no transfer of functional projections from L1 to L2, positing that learners transfer only lexical projections (VPs) during early acquisition. Higher-level functional projections (e.g., CP) are assumed to emerge later in development via the interaction of L2 input and principles of Universal Grammar (UG) without using L1 functional heads as templates. Most other models of transfer assume that functional projections from L1 transfer to L2, though they differ as to what this entails. Eubank's (1994a,b) *Weak Transfer* hypothesis holds that functional heads from L1 transfer along with their parameter settings (e.g., basic directionality), but L1-specific lexical feature-values associated with those heads do not transfer. *Full Transfer* models contend that L1 functional heads, their parameter settings, and their associated feature values serve as the initial interlanguage template for L2 development (e.g., Schwartz & Sprouse 1994, 1996).[2] As far as island-insensitivity in L1 Norwegian can be attributed to the presence of extra functional structure, observing comparable island insensitivity in L2 English would constitute evidence for transfer of that functional structure.

[1] A reviewer urged us to consider processing-based explanations for island effects as an alternative to grammatical approaches. We do not reject the idea that the acceptability of island constructions is influenced by processing factors, but we consider it unlikely that island effects are entirely reducible to processing. Cross-linguistic variation in island-sensitivity militates against such a simple reduction, insofar as cognitive resources are assumed not to vary systematically as a function of native language. The argument is even stronger when the two languages that exhibit different island-sensitivity share as many surface syntactic similarities as do English and Norwegian.

[2] We remain agnostic as to whether 'Full Transfer' entails *automatic wholesale* transfer of a whole grammar at the beginning of acquisition or whether individual units of functional structure can transfer on a *property-by-property* basis in response to input (e.g., Westergaard 2019).

Conservativity and Transfer

Language learners often encounter input that is compatible with two (or more) analyses differing in generative capacity: a restrictive analysis that closely fits the observed data and another more powerful analysis that generates both the observed data and additional unattested sentences. In such cases the strings generated by the first analysis represent a *subset* of the strings generated by the more powerful analysis (with respect to a given phenomenon).[3] When learners must choose between the two analyses, they face a version of the classic subset-superset problem (e.g., Berwick, 1985; Wexler & Manzini, 1987; White, 1989a; see Yuan 1997 and Judy & Rothman 2010 for cases in L2): should they choose the more, or less, restrictive analysis? What if they choose the less restrictive analysis and it turns out the be incorrect? If so, rejecting the superset analysis may be difficult since strings consistent with the subset analysis are equally consistent with the superset analysis.

Similar learnability considerations apply in L2 acquisition, where the problem is aggravated by the possibility of transfer: If the learner's L1 supports the superset analysis and the analysis is transferred to L2, the result is an overly permissive L2 grammar that generates both acceptable and unacceptable L2 forms. The case of Norwegian is arguably such an instance: if Norwegian learners transfer their L1 functional structure to their analysis of L2 English filler-gap dependencies, they would be able to generate acceptable long-distance dependencies in English, but also island-violating dependencies that should be unacceptable in English.

Prior work in L1 acquisition has argued that learners can avoid erroneous overgeneralization by adopting conservative learning strategies that prefer restrictive analyses (e.g. Snyder 2007; Westergaard 2014).[4] In principle, it is possible that transfer is also conservative: L2 learners could eschew transferring features that would potentially over-generate or avoid transferring typologically marked structures (e.g. Mazurkewich, 1984) without direct evidence for those features. Previous research has shown that L2 learners are less conservative than L1 learners, but these studies have not directly considered the role that transfer might play in these situations (e.g., Clahsen and Muysken, 1986; White, 1989b; Anderssen et al., 2018).

If Norwegians treat embedded questions as non-islands in L2 English, this would constitute evidence against conservative transfer.

Retraction and Restructuring after Transfer

L2 learners can undo transfer of an L1 feature (e.g. restructure) based on *positive evidence* of conflict between L2 input data and L1 analyses. Many models assume that *direct* positive evidence of conflict with the L1 analysis of phenomenon, P, is *required* for restructuring the L2 analysis of P (see, e.g., Schwartz & Sprouse 1996): metalinguistic-type negative evidence (e.g. correction) is believed not to be useful for prompting underlying grammatical restructuring (Schwartz 1993; White 2003). For some basic phenomena like determining head-directionality the relevant evidence is in abundance, so restructuring should happen quickly. As the similarity between or number of (surface) forms predicted by L1 and L2 increases, however, the possibility of direct conflict diminishes: the relevant positive data are scarce, if they exist at all. In the absence of (enough) conflicting data, inappropriately transferred features are expected to persist late into acquisition or become fossilized (Hawkins et al.

[3] Here we deliberately frame the subset-superset distinction in terms of strings or *weak generative capacity* rather than subset/superset grammars. This formulation is compatible with at least two possibilities: (i) that the two analyses under consideration map to grammars in a proper subset/superset relation, or (ii) that they map to *intersecting* grammars that can be distinguished by their generative capacities in other areas.

[4] There are cases where children appear to overgeneralize in non-conservative ways (see, e.g., Mazurkewich & White 1984; Pinker 1989; Ambridge et al. 2013).

1993; Schwartz & Sprouse 1996; Franceschina 2005; Lardiere 2007, Judy & Rothman 2010, a.o.). As such, instances where L2 surface forms are a subset of acceptable L1 forms represent paradigm cases where 'persistent' or fossilized transfer should obtain. Given that (most) acceptable English filler-gap dependencies are compatible with a transferred Norwegian analysis, we predict that Norwegians are likely to have restructured their L2 English grammars if transfer has occurred.

Past Work on Learnability of Islands/Movement

Before proceeding to our experiments, we briefly consider past work that investigated islands in L2 acquisition to highlight the difference from our research questions. Most prior studies were framed as tests of access to principles of Universal Grammar (UG) during L2 acquisition rather than transfer.

Some earlier experiments explored if L1 speakers of languages without overt *wh*-movement accept island-violating *wh*-movement in English (Johnson and Newport, 1991, Wolfe Quintero, 1992; Li, 1998; White and Genesee, 1996; White and Juffs, 1998; Martohardjono, 1993). For example: As part of a larger study, Martohardjono (1993) had L1 Chinese and L1 Indonesian participants judge sentences with long-distance wh-dependencies in their L2 English. Test sentences contained *wh*-dependencies that into five types of constituents that are islands in English: embedded questions (wh-islands), RCs, complex NPs, adjunct clauses, and sentential subjects. Martohardjono found that participants in both groups correctly rejected island violations on a non-trivial portion of trials,[5] which was taken as evidence for access to UG constraints on wh-movement during L2 acquisition (see Li, 1998; White & Juffs, 1998 for similar conclusions).

Other experiments have tested whether learners accept island-violating L2 filler-gap dependencies that correspond to unacceptable dependencies in their L1. Martohardjono (1993) again provides an example. Martohardjono asked L1 Italian participants to rate the same English sentences as the native Chinese and Indonesian participants in the experiment above. In Italian, *wh*-movement from all five of the constituents is unacceptable, just as in English (Rizzi, 1982; Sprouse et al., 2016). Martohardjono found that Italian participants rejected the test sentences at rates comparable to L1 English natives.[6] These results demonstrate that participants do not allow island-violating dependencies in their L2 if those dependencies are unacceptable in L1, an empirical conclusion that is also supported by the growing body of research on the real-time processing of islands in L2 (Omaki & Schulz, 2012; Felser, Cunnings, Batterham & Clahsen, 2012; Kim, Baek & Tremblay, 2015).

The results above do not directly address the limits of transfer because they are in principle compatible with transfer either having or not having occurred. If speakers of non-*wh*-movement initially transferred their L1 analysis of *wh*-dependencies to L2 English, observing overt movement dependencies would prompt them to restructure and generate a new analysis for the observed forms in the L2 input. If transfer did not occur, they would similarly base an analysis of English *wh*-dependencies on input forms. Judgments of English dependencies, then, would be based on their input-driven analyses. In the case of Italian, if participants conservatively learn the distribution of acceptable English dependencies from the L2 input alone or transfer their L1 analysis, they should reject island-violating *wh*-dependencies all the same.

[5] Participants' judgments of L2 island violations were often at, or close to, chance. Insofar as participants did not accept the island violations outright, the results were taken as evidence of UG influence. However, in the absence of information about how participants judged acceptable sentences, it is not clear how strongly to interpret the results.

[6] Italian *wh*-movement exhibits sensitivity to *wh*-islands, but Italian relative clause movement does not (Rizzi, 1982). It would therefore be possible to test another superset-subset configuration using RC-movement in Italian.

Unlike prior experiments, our work tests whether transfer occurs by testing cases where the dependencies allowed by L1 constitute a larger set than is allowed in L2. If transfer occurs, we expect 'unlearning' the L1 analysis should prove difficult because there is arguably little to no direct evidence that would contradict the transferred analysis. As such, we expect the transferred analysis to persist and to affect participant judgments even despite high otherwise proficiency in the L2 and significant time knowing the L2 well. In particular, we expect participants to accept unacceptable L2 forms generable under the L1 analysis. We tested whether L2 speakers of English make such errors with two acceptability judgment studies. To preview our main results, we find evidence for the predicted non-conservative transfer from L1 Norwegian to L2 English. However, we also find evidence that suggests some degree of restructuring: Norwegians do not uniformly treat embedded questions as non-islands in English as they do in Norwegian. We consider the implications of these facts in the General Discussion.

Experiments

We ran two acceptability judgment studies that tested Norwegian speakers' intuitions about the acceptability of relative clause dependencies in configurations like (3b) and (4b) in both English and Norwegian. We henceforth refer to such examples as *Wh-Trace Configurations* to highlight two characteristics of the constructions: (i) the islands in question are embedded questions (*wh*-islands) and (ii) the filler is associated with a subject gap/trace immediately adjacent to the embedded *wh*-word. Both aspects of the constructions are presumed to result in unacceptability in English: (i) because of an island violation, and (ii) because it is unacceptable in (most dialects of) English to have a gap next to an overt element in the complementizer domain (so called *Comp-trace effects*, Perlmutter, 1971; Chomsky & Lasnik, 1977). As both experiments had the same design, we present information about the materials, procedure, and analysis before discussing the specifics of each experiment.

Materials & Design

Both experiments employed the factorial definition of island effects developed by Sprouse (2007) and used in many recent studies of island-sensitivity cross-linguistically (e.g., Sprouse et al., 2011, Sprouse et al., 2012, and Sprouse et al., 2016; Kush, Lohndal & Sprouse, 2018, 2019).

The standard 2×2 factorial design for island effects crosses the factors STRUCTURE and DISTANCE. All test sentences contain a filler-gap dependency. DISTANCE controls the length of the dependency, while STRUCTURE controls the presence or absence of an island configuration. (5) illustrates the design with an item for testing Wh-Trace configurations.

(5) *Sample Wh-Trace Item (English Conditions)*
The sailors …

a. found *someone* that __ knew [that the signal meant danger].	SHORT | NO-ISLAND
b. saw *the signal* that they knew [___ meant danger].	LONG | NO-ISLAND
c. found *someone* that __ knew [**what** the signal meant].	SHORT | ISLAND
d. saw *the signal* that they knew [**what** ___ meant].	LONG | ISLAND

In (5) the filler-gap dependency is a relative clause dependency. DISTANCE determined whether the head of the RC (either *someone* or *the signal*) was linked to the highest subject position in the RC (Short) or to the embedded subject position (Long). STRUCTURE manipulated whether the most deeply embedded clause inside the relative clause was declarative (No-Island) or an embedded question (Island). The Long-Island condition corresponds to the only sentence with an 'island violation'. According to the logic of the factorial design, an island effect is defined as a STRUCTURE × DISTANCE

interaction: it reflects the residual unacceptability associated with a structure like (5d) once the independent effects that long-distance extraction and structural complexity have on acceptability have been accounted for.

We crossed the standard 2×2 manipulation above with an additional factor: LANGUAGE. Norwegian counterparts for all English items were created, (6), resulting in a 2×2×2 design.

(6) *Sample Wh-Trace Item (Norwegian Conditions)*
Sjømennene …
Sailors.DEF.PL

a. fant *noen* som __ visste [at signalet betydde fare]. SHORT | NO-ISLAND
found someone that knew that signal.DEF meant danger

b. så *signalet* som de visste [at __ betydde fare]. LONG | NO-ISLAND
saw *signal* that they knew at ___ meant danger

c. fant *noen* som __ visste [**hva** signalet betydde]. SHORT | ISLAND
found someone that knew what signal.DEF meant

d. så *signalet* som de visste [**hva** __ betydde]. LONG | ISLAND
saw *signal* that they knew what ___ meant

In addition to Wh-Trace items, we tested sensitivity to another island type: Subject islands. The islandhood of subject phrases is determined by different syntactic constraints (e.g. the *Condition on Extraction Domains* of Huang 1982) than the embedded questions. As a result, the extra functional structure that permits Wh-Trace island violations should have no effect on the islandhood of subjects. Thus, subjects should be islands in both Norwegian and English. This prediction has been verified by previous studies using the factorial design (Sprouse et al., 2011; 2016 and Kush et al., 2018, 2019).

We adapted materials from Kush et al. (2018, 2019) to test the acceptability of RC-dependencies into subject islands. The design crossed STRUCTURE × DISTANCE × LANGUAGE, yielding 8 conditions as exemplified below.

(7) *Sample Subject Island Item (English Conditions)*
The judge …

a. met the lawyer that __ hoped that the report would confirm the suspicions. SHORT | NO-ISLAND

b. read the report that the lawyer hoped would __ confirm the suspicions. LONG | NO-ISLAND

c. met the lawyer that __ hoped that the information in the report would confirm the suspicions. SHORT | ISLAND

d. read the report that the lawyer hoped that [the information in __] would confirm the suspicions. LONG | ISLAND

(8) *Sample Subject Island Item (Norwegian Conditions)*
Dommeren…

a. møtte advokaten som __håpet at rapporten ville bekrefte mistankene. SHORT | NO-ISLAND

b. leste rapporten som advokaten håpet at __ ville bekrefte mistankene. — LONG | NO-ISLAND

c. møtte advokaten som __ håpet at opplysningene i rapporten ville bekrefte mistankene. — SHORT | ISLAND

d. leste rapporten som advokaten håpet at [opplysningene i __] ville bekrefte mistankene. — LONG | ISLAND

Subject island judgments provide an independent baseline of island-sensitivity that is not expected to be affected by the hypothesized transfer of functional structure.

Procedure

Test items were distributed across lists according to a Latin Square design and intermixed among filler sentences. The experiment was hosted on IbexFarm (Drummond, 2012). Participants participated on their own personal computers. Sentences were presented one at a time. Participants rated their acceptability on a 7-point scale. All participants rated English items first before judging a Norwegian block to minimize L1 interference. Instructions were presented in English and participants received a break between English and Norwegian blocks.

Analysis

Raw ratings were z-score transformed before analysis. We z-scored ratings by participant and language tested. Z-scoring by-participant helps to control for biases in in how individual participants used the 7-point scale. Z-scoring by-language for each participant helps control for the fact that participants may use the scale differently in their L1 and L2 (Sorace, 1996; Spinner and Gass, 2019).

Z-scores were analyzed with linear mixed effects models implemented using the lme4 (Bates, Maechler, Bolker & Walker, 2015) and lmerTest (Kuznetsova, Brockhoff & Christensen, 2017) packages in R (R Core Team, 2013). All models included fixed effects of STRUCTURE, DISTANCE and their interaction and random intercepts for both subject and item. When appropriate we also included fixed effects of ISLAND TYPE and LANGUAGE. We included by-subject random slopes for STRUCTURE, DISTANCE and their interaction when such models converged. When the more complex model did not converge, we simplified the random effects structure. Details of individual models are presented in the tables of statistical results below. P-values were computed using likelihood ratio tests. Effect size was measured as a Difference-in-differences (DD) score (Maxwell & Delany, 2003). DD scores were calculated by-participant.

Experiment 1

Materials

16 items of 8 conditions apiece were created for each island type following the DISTANCE × STRUCTURE × LANGUAGE design. The 32 test items (16 Wh-Trace, 16 Subject Island) were interspersed among 76 filler sentences (38 English, 38 Norwegian). Each set of language-specific filler sentences contained 22 unacceptable and 16 acceptable fillers varying in length and complexity.

Experiment 1a: Native English Controls

31 native English volunteers recruited as control participants via social media (mean age = 38.0, sd=11.9, 17 female; 27 from the United States) judged sentences in the English block of the experiment on Ibex Farm. One participant was excluded from analysis for having multiple response times <500ms. Because participants rated only the English sentences, participants rated 4 tokens per condition per island.

Results

Average acceptability judgments by condition are found in Figure 1. A summary of statistical analysis is found in Table 1.

[[INSERT FIGURE 1 ABOUT HERE]]

Effect	Beta (sd)	*t*	*p*
DISTANCE	**1.29 (0.09)**	**14.291**	**< .000**
STRUCTURE	**1.63 (0.09)**	**18.404**	**< .000**
LANGUAGE	-0.233 (0.13)	-1.835	0.071
DISTANCE × STRUCTURE	**-1.55 (0.12)**	**-13.351**	**< .000**
DISTANCE × ISLAND	0.230 (0.12)	1.962	0.050
STRUCTURE × ISLAND	0.0027 (0.12)	0.023	0.981
DIST × STRUCT × ISLAND	-0.0540 (0.16)	-0.329	0.742

Table 1. Summary of statistical analysis of native English Control judgments from Experiment 1a. Significant effects are in bold face. Model: zscore ~ DISTANCE * STRUCTURE * ISLAND + (1 + DISTANCE + STRUCTURE | subject) + (1|item)

Native English speakers rated RC dependencies into both subject phrases and embedded Wh-Trace constructions much lower than RC dependencies into non-islands. There were clear island effects (DISTANCE × STRUCTURE $p < .001$). The sizes of Subject and Wh-Trace island effects (DDs = 1.55, 1.69, respectively) did not differ significantly, as evidenced by the absence of a three-way DISTANCE × STRUCTURE × ISLAND TYPE interaction.

We also inspected by-participant ratings of the Subject and Wh-Trace *Long-Island* sentences to check for inter-trial consistency. Native English participants rejected Wh-Trace *Long-Island* sentences nearly uniformly. 28 of 30 participants rejected 4 out of 4 Wh-Trace *Long-Island* tokens, judging all below $z=0$. The two remaining participants rated a single Wh-Trace Long-Island token above $z=0$, but rejected the remaining 3 tokens.

Judgments of Subject Long-Island sentences showed slightly more variability. 14 participants rejected all four tokens that they rated. 12 participants rejected three of four tokens. Three participants exhibited more variability: two of the three rejected only two of four Subject Long-Island sentences and one participant rejected only one of four. Overall, however, participants rejected RC-dependencies into subject phrases on the clear majority of trials.

Experiment 1b: Norwegian L1, English L2

Participants

27 native speakers of Norwegian took part (16 female). Two participants' data were excluded because the participants reported exposure to English during infancy. Participants were students enrolled in the English Studies program at the Norwegian University of Science and Technology (NTNU) either at the bachelor's or master's level. Norwegian university students are assumed to have a proficiency in English at least commensurate to the Common European Framework of Reference for Languages (CEFR) at B2 level, as this is the minimum standard for enrollment for foreign students (see e.g., Samordna opptak, 2019). Participants filled out a short survey on their language background and their English exposure. An overview of responses to this survey is in Table 2.

	Mean (sd)	**Median**	**Range**
Age	23.1 (4.3)	22	19-39
Age When Began Learning English	6.56 (1.4)	6	5-10
English Spoken (hrs./week)	1.62 (1.7)	1	0-6
English Media (hrs./week)	5.14 (2.4)	4.5	1.5-12

English Proficiency (self-reported, 7pt scale)	5.6 (0.8)	6	4-7

Table 2. Demographic information for Norwegian participants in Experiment 1b.

Unlike the English control participants, Norwegian participants rated 8 items per island in each language (2 tokens per condition per language). Average judgments by language and island type are plotted in Figure 2.

[[INSERT FIGURE 2 ABOUT HERE]]

We first report the results of the omnibus DISTANCE × STRUCTURE × ISLAND TYPE × LANGUAGE analysis before planned analyses of each island type in isolation.

	Beta (sd)	*t*	*p*
DISTANCE	**0.630 (0.07)**	**9.024**	**< .000**
STRUCTURE	**0.743 (0.07)**	**10.664**	**< .000**
ISLAND TYPE	0.160 (0.11)	1.474	0.151
LANGUAGE	-0.010 (0.07)	-0.142	0.887
DISTANCE × STRUCTURE	**-1.313 (0.14)**	**-9.428**	**< .000**
DISTANCE × ISLAND TYPE	-0.181 (0.10)	-1.836	0.067
STRUCTURE × ISLAND TYPE	**-0.692 (0.10)**	**-7.030**	**< .000**
DISTANCE × LANGUAGE	0.180 (0.14)	1.296	0.195
STRUCTURE × LANGUAGE	-0.134 (0.14)	-0.965	0.335
ISLAND TYPE × LANGUAGE	**0.279 (0.10)**	**2.837**	**0.005**
DIST × STRUCT × ISLAND	**0.970 (0.20)**	**4.925**	**< .000**
DIST × STRUCT × LANGUAGE	**0.576 (0.28)**	**2.073**	**0.038**
DIST × ISLAND × LANGUAGE	-0.230 (0.20)	-1.170	0.242
STRUCT × ISLAND × LANGUAGE	-0.170 (0.20)	-0.864	0.388
DIST × STRUCT × ISLAND × LANG	0.223 (0.39)	0.569	0.569

Table 3. Omnibus statistical analysis of judgments from Experiment 1b. Significant effects are in boldface. Model: zscore ~ DISTANCE * STRUCTURE * ISLAND * LANGUAGE + (1|subject) + (1|item).

Sentences with long-distance RC dependencies were rated lower on average than sentences with short RC dependencies ($p < .000$), as were sentences with island structures ($p < .000$). However, these main effects were qualified by several higher-order interactions. We focus on the two three-way interactions. The significant interaction of DISTANCE × STRUCTURE × LANGUAGE ($p = .0385$) reflects the fact that Norwegians exhibited larger average DISTANCE × STRUCTURE island effects in English than in Norwegian. The significant DISTANCE × STRUCTURE × ISLAND TYPE interaction ($p < .000$) indicates that Norwegians exhibited larger DISTANCE × STRUCTURE island effects for Subject island sentences than for Wh-Trace sentences. Although we did not observe a significant four-way interaction, we conducted planned comparisons of the three-way interaction of DISTANCE × STRUCTURE × LANGUAGE for each island type separately.

Subject Islands. A statistical summary is given in Table 4. The size of Subject island effect was significantly larger in English (DD = 1.69) than in Norwegian (DD = 1.01), as indicated by a DISTANCE × STRUCTURE × LANGUAGE interaction ($p = .016$). Resolving this interaction confirmed that significant, sizable subject island effects were present both in English and Norwegian (ts = -10.05, -6.16, respectively; ps < .001).

	Beta (sd)	*t*	*p*
DISTANCE	**0.627 (0.06)**	**10.241**	**< .000**
STRUCTURE	**0.737 (0.09)**	**9.003**	**< .000**
LANGUAGE	-0.010 (0.06)	-0.155	0.877

DISTANCE × STRUCTURE	**-1.321 (0.12)**	**11.107**	**< .000**
DISTANCE × LANGUAGE	0.180 (0.12)	1.521	0.129
STRUCTURE × LANGUAGE	-0.134 (0.13)	-1.030	0.308
DIST × STRUCT × LANGUAGE	**0.576 (0.24)**	**2.426**	**0.016**

Table 4. Statistical analysis of judgments of the subject island items from Experiment 1b. Significant effects are in boldface. Model: zscore ~ DISTANCE * STRUCTURE * LANGUAGE + (1 + DIST + STRUCT + LANG |subject) + (1|item).

Wh-Trace Islands. A statistical summary can be found in Table 5. Again, the three-way DISTANCE × STRUCTURE × LANGUAGE interaction was significant ($p < .01$). Resolving the three-way interaction revealed that although there was a significant Wh-Trace island effect in English ($t = -4.10$, $p < .001$; DD = 0.38), there was not a significant WhTrace island effect in Norwegian (DD = -.01). Visual inspection of Figure1 confirms the absence of even trend towards an interaction in Norwegian.

	Beta (sd)	*t*	*p*
DISTANCE	**0.441 (0.08)**	**5.333**	**< .000**
STRUCTURE	0.051 (0.08)	0.661	0.510
LANGUAGE	**0.266 (0.07)**	**3.556**	**< .001**
DISTANCE × STRUCTURE	**-0.338 (0.15)**	**-2.252**	**0.025**
DISTANCE × LANGUAGE	-0.048 (0.15)	-0.317	0.751
STRUCTURE × LANGUAGE	-0.302 (0.17)	-1.732	0.095
DIST × STRUCT × LANGUAGE	**0.797 (0.30)**	**2.661**	**0.008**

Table 5. Statistical analysis of judgments of the wh-trace island items from Experiment 1b. Significant effects are in boldface. Model: zscore ~ DISTANCE * STRUCTURE * LANGUAGE + (1 + DIST + STRUCT + LANG |subject) + (1|item).

The Wh-Trace island effect observed in English is smaller in magnitude than the subject island effects in either language, while the average rating of the English *Wh-Trace Long-Island* sentence is considerably higher (roughly -0.25) than the English *Subject Long-Island* sentence (roughly -0.80). Following Kush, Lohndal & Sprouse (2018, 2019), we investigated whether the smaller effect reflected *inconsistent* judgments across trials. Figure 3 plots the distribution of z-scores in both *Long* conditions for each island-language combination. Response consistency is reflected in the degree to which judgments in a condition follow a unimodal distribution. Inconsistent judgments manifest as bimodal or uniform distributions. In each of the island-language pairs, the *Long-NoIsland* condition provides a baseline level of consistency against which to judge the responses in the *Long-Island* conditions. The extent of the overlap between the *Long-NoIsland* and *Long-Island* judgments provides a rough way of approximating the extent to which the RC-dependencies into islands were perceived as run-of-the-mill long-distance dependencies.

[[INSERT FIGURE 3 ABOUT HERE]]

Beginning with subject island sentences, we observe relatively little overlap between that the ratings for *Long-Island* and *Long-NoIsland* sentences. Judgments of Subject *Long-NoIsland* sentences cluster unimodally around the higher end of the scale ($z = +1$), with a thicker left tail. Judgments of the Subject *Long-NoIsland* condition, by contrast, cluster at the opposite end of the scale ($z = -1$) and exhibit less of a right skew.

Judgments of *Long* conditions in the Subject island sub-experiment provide a template for consistent judgment. Judgments in the Wh-Trace sub-experiments clearly do not conform to that template. Judgments in the Norwegian Wh-Trace *Long-Island* condition are consistent with general acceptability: z-scores are unimodally distributed about the high end of the scale with a fat left tail. The pattern of responses indicates that participants perceived the test sentences as unobjectionable on

most trials. Bimodality in the corresponding *Long-NoIsland* condition suggests that participants were less consistent in judging those sentences. Turning to the English Wh-Trace sub-experiment we see bimodality in both *Long* conditions, though the larger mode falls on opposite ends of the range between conditions. Norwegian participants tended to accept *Long-NoIsland* sentences more often than reject them, but there were still a number of trials where they judged the sentences to be unacceptable. Most relevant to our purposes, the Norwegian participants often rejected *Long-Island* sentences, but there was a non-negligible number of trials on which they accepted structures that native English speakers reject.

Individual Differences

Analysis of the rating distributions shows that there was inter-trial inconsistency in the ratings of English Wh-Trace *Long-Island* sentences, but it does not establish whether the cause was inter- or intra-participant inconsistency. To ascertain whether individual participants were inconsistent, we plotted each participant's maximum judgment against their minimum judgment for each island-language combination (see Kush et al., 2019). In Figure 4, each dot corresponds to an individual participant.

[[INSERT FIGURE 4 ABOUT HERE]]

For the purposes of the analysis we adopt a crude definition of 'acceptance' and 'rejection': we treat all judgments that fall below $z = 0$ as *rejections* and all judgments that are above $z = 0$ as *acceptances*. Using this coarse categorization technique permits identification of three participant response types: Participants that rejected both tokens of an island type occupy quadrant 3 (bottom left). Those that accepted both tokens occupy quadrant 1 (top right). Those that occupy quadrant 4 (top left) rated island tokens inconsistently, accepting one and rejecting the other.

A few participants accepted one or both Norwegian subject island tokens, but subject island judgments otherwise exemplify consistent rejection: In Panels 1 and 2 of Figure 4 most participants fall into quadrant 4. Judgments of the Norwegian Wh-Trace sentences show a different pattern: all participants fell into quadrant 1 (consistent accepters) or quadrant 4 (inconsistent raters). Judgments of English Wh-Trace islands show more variability. 11 participants consistently rejected Wh-Trace islands in English and 3 consistently accepted the constructions. The remaining 11 participants rated the sentences inconsistently. This level of inter- and intra-participant inconsistency stands in contrast to the relative uniformity of the same participants' judgments of English Subject island tokens.

Given the differences in participant response patterns for the English Wh-Trace, we conducted an exploratory analysis of whether individual variability correlated with self-reported proficiency, weekly hours of English spoken, or English media consumption. We used participants' English Wh-Trace island DD score as the dependent measure of island sensitivity. A positive correlation between DD score and individual measure would be expected on the assumption that increased exposure or proficiency made participants behave more like native English speakers. Hours of spoken English did not correlate with DD score ($|t| < 1$), nor did self-reported English proficiency ($|t| < 1$). There was a small, but significant negative correlation between DD score and English media consumption ($t = -2.036$, $p < .05$; adjusted $R2 = .116$). As Figure 5 shows, this correlation indicates – counter-intuitively – that participants who consumed more English media showed reduced sensitivity to English Wh-Trace island effects.

[[INSERT FIGURE 5 ABOUT HERE]]

Discussion

As expected, English participants rejected RC-dependencies into subject phrases. Norwegian participants also rejected subject island violations in their L1 Norwegian and L2 English. These

results are expected, if subjects are islands in both languages and the extra functional structure that allows filler-gap dependencies into embedded questions does not amnesty subject island violations.

Participants diverged in their judgments of Wh-Trace items. Native English speakers exhibited large Wh-Trace island effects, rejecting RC-dependencies in Wh-Trace configurations. We failed to find a Wh-Trace island effect in Norwegian. Norwegian participants generally accepted RC-dependencies into Wh-Trace configurations in their L1 as readily as RC-dependencies into declarative complement clauses. Interestingly, we found a significant island effect with English Wh-Trace constructions, indicating that Norwegians rated RC-dependencies in Wh-Trace constructions less acceptable on average than RC-dependencies into non-island declarative complement clauses. However, the island effect was smaller than subject island effects, because Norwegian participants rated English Wh-Trace islands *inconsistently*: participant ratings were a mix of 'accept' and 'reject' trials. We defer further discussion and interpretation of this finding to the General Discussion.

The number of trials where Norwegian participants accepted English Wh-Trace violations provides suggestive support for transfer from L1 Norwegian to L2 English. However, the experiment was relatively low-powered, with only two observations of the relevant configuration per participant. We wished to test if our findings would replicate, and whether participants would provide more consistent judgments of Wh-Trace island violations in English if given more trials. Therefore, we ran Experiment 2, in which we doubled the number of observations per participant. We also increased our sample size and drew from a wider pool.

Experiment 2

Participants

Forty-nine native speakers of Norwegian took part in experiment 2 (29 female). Like the participants in Experiment 1, participants in Experiment 2 were enrolled as bachelor's and master's students in a Norwegian university. Unlike the previous participants, participants in Experiment 2 were enrolled in a wide range of degree programs, not only English. All these courses of study presuppose that students have studied English from upper secondary school and have achieved minimum proficiency at CEFR B2 level. Participants provided the same information as in Experiment 1. Table 6 provides an overview of descriptive statistics.

	Mean (sd)	**Median**	**Range**
Age	23.3 (3.9)	23	19-39
Age When Began Learning English	6.58 (1.7)	6	5-10
English Spoken (hrs./week)	1.35 (1.4)	1	0-6
English Media (hrs./week)	4.9 (2.6)	4	1.5-12
English Proficiency (self-reported)	5.7 (0.7)	6	4-7

Table 6. Demographic information for Norwegian participants in Experiment 2.

Materials

Participants rated the same items as in Experiment 1 plus 16 new Wh-Trace items. As a result, participants judged 4 tokens per condition per language in the Wh-Trace sub-experiment instead of 2 as in Experiment 1.

Results

Participants' average judgments by island type and language are plotted in Figure 6. A summary of the omnibus statistical analysis can be found in Table 7. As in the analysis of Experiment 1, we focus only on the highest-order interaction effects.

[[INSERT FIGURE 6 ABOUT HERE]]

	Beta (sd)	*t*	*p*
DISTANCE	**0.559 (0.03)**	**18.986**	**< .000**
STRUCTURE	**0.390 (0.03)**	**13.349**	**< .000**
ISLAND TYPE	**0.262 (0.07)**	**3.685**	**< .001**
LANGUAGE	**0.176 (0.03)**	**5.984**	**< .000**
DISTANCE × STRUCTURE	**-0.720 (0.06)**	**12.305**	**< .000**
DISTANCE × ISLAND TYPE	**-0.453 (0.06)**	**-7.695**	**< .000**
STRUCTURE × ISLAND TYPE	**-0.632 (0.06)**	**10.795**	**< .000**
DISTANCE × LANGUAGE	**0.167 (0.06)**	**2.862**	**.004**
STRUCTURE × LANGUAGE	**-0.272 (0.06)**	**-4.634**	**< .000**
ISLAND TYPE × LANGUAGE	0.092 (0.06)	1.575	0.115
DIST × STRUCT × ISLAND	**0.946 (0.12)**	**8.082**	**< .000**
DIST × STRUCT × LANGUAGE	**0.627 (0.12)**	**5.360**	**< .000**
DIST × ISLAND × LANGUAGE	-0.051 (0.12)	-0.436	0.663
STRUCT × ISLAND × LANGUAGE	-0.150 (0.12)	-1.278	0.201
DIST × STRUCT × ISLAND × LANG	0.066 (0.23)	0.283	0.778

Table 7. Omnibus statistical analysis from Experiment 2. Significant effects are in boldface. Model: zscore ~ DISTANCE * STRUCTURE * ISLAND * LANGUAGE + (1|subject) + (1|item)

There was a significant DISTANCE × STRUCTURE × ISLAND interaction ($p < .000$): Subject island effects were, on average, larger than Wh-Trace island effects, irrespective of language. The significant DISTANCE × STRUCTURE × LANGUAGE interaction reflects that when collapsing across Subject and Wh-Trace island sentences participants exhibited numerically smaller average island effects in their judgment of Norwegian test items than with English test items. We proceed to the planned comparisons, focusing on each island type separately.

Subject Islands. A statistical summary is in Table 8. As in Experiment 1, *Long* sentences were rated lower on average than *Short* sentences ($p < .000$) and *Island* sentences were rated lower *NoIsland* sentences ($p < .000$) collapsing across languages. The three-way DISTANCE × STRUCTURE × LANGUAGE interaction was again significant ($p < .001$). The interaction was driven by the fact that the Norwegian Subject Island effect was numerically smaller than the English effect. However, subject island effects were large in both English and Norwegian (DDs = 1.47, 0.89, respectively) and significant (DISTANCE × STRUCTURE: ts = -12.54, -8.01, respectively; ps < .000).

	Beta (sd)	*t*	*p*
DISTANCE	**0.790 (0.05)**	**16.076**	**< .000**
STRUCTURE	**0.712 (0.06)**	**12.892**	**< .000**
LANGUAGE	**0.136 (0.04)**	**3.160**	**.002**
DISTANCE × STRUCTURE	**-1.179 (0.08)**	**-14.370**	**< .000**
DISTANCE × LANGUAGE	**0.196 (0.08)**	**2.385**	**.017**
STRUCTURE × LANGUAGE	**-0.194 (0.08)**	**-2.363**	**.018**
DIST × STRUCT × LANGUAGE	**0.588 (0.16)**	**3.583**	**< .001**

Table 8. Statistical analysis of judgments of the subject island items from Experiment 2. Significant effects in boldface. Model: zscore ~ DISTANCE * STRUCTURE * LANGUAGE + (1 + DIST + STRUCT + LANG |subject) + (1|item)

Wh-Trace Islands. Table 9 presents a statistical summary. *Long* conditions were rated significantly lower on average than *Short* conditions ($p < .000$) and *Island* conditions lower than *NoIsland* conditions ($p < .05$). The DISTANCE × STRUCTURE × LANGUAGE interaction was significant ($p < .001$). Figure 6 makes evident that the interaction is because Norwegians exhibited an average island

effect for English sentences (DD = 0.568), but not for Norwegian sentences (DD = -0.069). Follow-up analyses verified that there was a significant Wh-Trace island effect in English ($t = -4.83$, $p < .001$), but not in Norwegian ($t < 1$).

	Beta (sd)	*t*	*p*
DISTANCE	**0.332 (0.04)**	**7.995**	**< .000**
STRUCTURE	**0.075 (0.04)**	**2.135**	**0.034**
LANGUAGE	**0.221 (0.04)**	**5.810**	**< .000**
DISTANCE × STRUCTURE	**-0.247 (0.07)**	**-3.499**	**0.001**
DISTANCE × LANGUAGE	**0.142 (0.07)**	**2.034**	**0.042**
STRUCTURE × LANGUAGE	**-0.349 (0.07)**	**-4.988**	**< .000**
DIST × STRUCT × LANGUAGE	**0.662 (0.14)**	**4.753**	**< .000**

Table 9. Statistical analysis of judgments of the wh-trace island items from Experiment 2. Significant effects in boldface. Model: zscore ~ DISTANCE * STRUCTURE * LANGUAGE + (1 + DIST * STRUCT + LANG |subject) + (1|item)

We again examined the distribution of ratings in *Long* conditions for all island-language pairs. Distributions are plotted in Figure 7. Subject *Long-Island* and *Long-NoIsland* distributions are roughly bimodal with modes at opposite ends of the rating scale. However, in the Norwegian Wh-Trace sentences, the distribution of ratings for *Long-Island* sentences is essentially indistinguishable from *Long-NoIsland* sentences: participants accepted the majority of test sentences in both conditions. English Wh-Trace judgments diverged from the ratings of their Norwegian counterpart sentences. Participants generally accepted *Long-NoIsland* sentences, judgments of *Long-Island* sentences are bimodally distributed. The group as a whole appears to accept and reject English Wh-Trace sentences with near equal frequency.

[[INSERT FIGURE 7 ABOUT HERE]]

Figure 8 plots individuals' minimum and maximum ratings for each island-language pair, to visualize rating consistency.

[[INSERT FIGURE 8 ABOUT HERE]]

Participants consistently rejected Subject Island violations in English and were generally consistent in their judgment of Norwegian Subject Island violations, as evidenced by the clustering in quadrant 3 in panels 1 and 2 of Figure 8.

Panel 3 shows that every participant accepted at least one Norwegian Wh-Trace island token– most participants fall into quadrant 4 – while many accepted all four tokens. Panel 4 indicates that almost all participants accepted at least one English Wh-Trace island token.

Figure 8 only provides information about the range of individual participants' judgments. We were also interested in how many of the 4 *Long-Island* Wh-Trace tokens each participant accepted. Therefore, we binned participants by how many tokens they rated above 0. The result is in Table 10. 40 of 49 participants accepted 3 or 4 Norwegian Wh-Trace island tokens and none rejected all 4 tokens. Judgments of English Wh-Trace tokens showed less consistency – fewer participants accepted most of the items (18 of 49). Five participants consistently rejected *Long-Island* tokens. Nevertheless, Norwegian participants clearly displayed a different response pattern than Native English speakers in Experiment 1a, where all participants either uniformly rejected the Wh-Trace island tokens, or rejected 3 of 4.

Number of tokens	Norwegian	English

rated $z > 0$	Wh-Trace Items	Wh-Trace Items
0	0	5
1	1	14
2	8	12
3	23	10
4	17	8

Table 10. Participants binned by the number of Wh-Trace island violation items they rated above the midpoint of the scale

One question that Table 10 leaves unaddressed is how strongly participants' judgments in the Norwegian Wh-Trace experiment correlate with their judgments in English. We addressed this question in two follow-up analyses. First, we plotted participants' Wh-Trace DD scores in Norwegian against their DD scores in English, to determine whether there was a correlation between island effect size. This plot is in Figure 9a. Second, we looked for a correlation between individual participants' probability of accepting a Wh-Trace island violation in Norwegian and English. The correlation plot is provided in Figure 9b.

[[INSERT FIGURE 9 ABOUT HERE]]

There was no reliable correlation between Norwegian and English Wh-Trace DD scores. As Figure 9a makes apparent, there were many participants who exhibited no island effects in Norwegian ($z \leq 0$), but nevertheless had a positive Wh-Trace DD score in English. Figure 9b shows a numeric trend such that participants who accepted a high proportion of Wh-Trace island violations in Norwegian were slightly more likely to accept Wh-Trace island violations in English, though this correlation was not significant (Adjusted $R^2 = .015$; $t = 1.32$). The correlation was weakened by the group of 19 participants who readily accepted Wh-Trace island violations in Norwegian (> 50%), but were less likely to do so in English. Importantly, all but five of these participants still accepted Wh-Trace violations in English. Finally, we checked whether any of the three individual-level variables correlate with a participant's Wh-Trace DD score. None of the measures correlated with DD score ($ts < 1$).

General Discussion

Embedded questions are islands for filler-gap dependency creation in English, but not in Norwegian. The difference between the two languages has been linked to extra functional structure in the left-periphery of the Norwegian clause (Vikner et al. 2017; Kush et al. 2018, 2019). We were interested in determining whether Norwegians transfer this extra functional structure from their L1 to their L2, English. We reasoned that if Norwegians transfer the functional structure to English, they should erroneously treat embedded questions as non-islands in English. Insofar as the set of acceptable Norwegian filler-gap dependencies represents a superset of the acceptable English filler-gap dependencies, acquiring the appropriate generalization in English should prove difficult if transfer has occurred. The difficulty reflects the fact that there is arguably little if any direct evidence to counter-exemplify the less restrictive hypothesis (White, 1989a).

To test whether such transfer occurs, we tested whether adult Norwegian speakers accept filler-gap dependencies into embedded questions (*wh*-islands) in English. Our results provide evidence of transfer from L1 Norwegian L2 English: Participants accept filler-gap dependencies into *wh*-islands in English even though they have never encountered those structures in their English input. Importantly, participants do not accept all island-violating filler-gap dependencies in L2 English, as evidenced by participants' consistent rejection of subject-island-violating filler-gap dependencies. The fact that subject island violations were consistently rejected militates against an interpretation that attributes Norwegian participants' acceptance of English Wh-Trace island violations to general

island-insensitivity in L2. As predicted by transfer, our participants only accepted island-violations in English if the corresponding dependency was acceptable in Norwegian.

Insofar as the non-island status of embedded questions is due to extra CP-level functional structure, our results are consistent with models that permit such transfer, including *Weak Transfer* (Eubank 1994a,b) or traditional *Full Transfer* models (Schwartz & Sprouse, 1994, 1996) over models that restrict transfer to minimal grammatical information (Vainikka & Young-Scholten, 1994, 1996).

We predicted that transfer of L1 functional structure could lead native Norwegians into a 'superset trap': having assumed that an analysis that allows more filler-gap dependencies than are acceptable in English, learners would be unable to retract to a more restrictive analysis. All else equal, we would therefore predict that Norwegian participants should accept island-violating dependencies as often in L2 English as they do in L1 Norwegian. Participant judgments yielded a more complicated picture: Almost all participants accepted Wh-Trace island violations in English on some portion of trials, but roughly one-third of our participants accepted the English structures less readily than in Norwegian.

The Source of Inconsistent Judgments

Participants' inconsistent judgments of English *wh*-island violations are consistent with two broad interpretations. First, participants may have rejected the sentences simply due to their increased complexity. This could occur if participants have greater difficulty processing wh-dependencies in their L2 than in their L1 (Juffs & Harrington, 1995; Juffs 2005). We point out that this explanation presupposes that transfer must have occurred, otherwise the Norwegians would not accept the island-violations in English at all. The explanation holds, however, that Norwegians' tendency to probabilistically reject island-violations does *not* constitute evidence of learning the appropriate English analysis: rejection occurs for orthogonal, extra-grammatical reasons.

The second option, which we favor, is that probabilistic rejection provides evidence of learning and partial restructuring. By 'restructuring' we simply mean that changes are made to some aspect of the holistic system or feature set transferred from L1. These changes could represent *target-like restructuring*, such that Norwegian speakers specifically reject the extra functional structure from Norwegian and adopt a simplified left-periphery identical to native English speakers. Alternatively, Norwegians could engage in ad-hoc *compensatory restructuring* wherein other grammatical changes are made to ensure closer surface alignment with acceptable English forms, without directly retracting the L1 functional structure. Our current results do not allow us to distinguish these two possibilities. Which of these outcomes is more likely depends, in part, on what types of L2 input learners receive as evidence that the set of filler-gap dependencies is different in English and how directly that evidence contradicts the L1 analysis. We consider the issue of evidence in the input presently.

Participants' stochastic or inconsistent judgments are compatible with the notion that they have learned that the distribution of filler-gap dependencies differs between the two languages, but that learning or restructuring is not 'complete'. Within a parameter-setting model of L2 acquisition (Schwartz & Sprouse, 1996; White, 2003) or a grammar competition/multiple grammars model (Amaral & Roeper, 2014; Rankin 2014), this uncertainty could be modeled as a probabilistic competition between different grammars. Transfer would entail that Norwegian learners begin acquisition by assigning a high probability to their L1 analysis. Over time, however, they would accumulate evidence against that analysis and would shift probability to a more restrictive analysis (provided they could avoid the preemption problem; Trahey & White 1993; Rothman & Iverson 2013).

Evidence of Difference

The question remains what cues learners could use in their English input as evidence in favor of the restrictive analysis. *Negative evidence* could, in principle, play a role. We consider *direct* negative

evidence in the form of corrections an implausible mechanism given: (i) the relative infrequency of relevant productions, (ii) the unreliability and ambiguity of interlocutors' correction, (iii) the low probability that *wh*-island violations are ever addressed explicitly in the English classroom (see, Carroll, 1995; 2001; Schwartz 1993). *Indirect* negative evidence is another option: if Norwegian learners of English expect to encounter English *wh*-island violations at a rate comparable to Norwegian, then the absence of the structures could over time lead the learner towards the restrictive hypothesis. Prior research has argued that indirect evidence may play a role in L1 (Rohde & Plaut, 1999, Regier & Gahl, 2004, Foraker et al., 2009, Perfors, Tenenbaum, & Regier, 2011, Ramscar et al., 2013) and L2 acquisition (Plough, 1994; Dahl, 2004). However, it is unclear whether the frequency of the island-violations is high enough in L1 to form the basis for strong predictions in L2.

Direct positive evidence of the unacceptability of *wh*-island violations does not occur, but some learning models allow *indirect positive evidence* to play a role (e.g., Pearl & Mis, 2016 or more traditional parameter-based models). Learners can rely on indirect positive evidence if there exist implicational relations between observed (non-island) structures and the possibility of island violations. Under the assumption that additional CP-level functional structure underlies the non-island status of embedded questions in Norwegian, Norwegians would require evidence that this structure is absent in English. Such evidence is only possible if some overt property of the English CP-domain conflicts with the Norwegian analysis. What could such cues be?

We assume that most sentences do not provide unambiguous evidence for deep differences in functional structure of the CP domain, given the similarity of surface word order patterns in the two languages. However, one piece of evidence might prove useful:

It has been suggested that evidence for an articulated CP-domain in Mainland Scandinavian comes from *embedded V2* phenomena (e.g. Vikner et al. 2017). Mainland Scandinavian languages exhibit V2 word order in main clauses (9a; Holmberg & Platzack 1995): the finite verb (*skal*) is the second constituent in linearly regardless of whether a subject (9a) or non-subject (9b) occupies sentence-initial position:

(9) a. Han **skal** antakeligvis ikke synge i-morgen.
He shall presumably NEG sing tomorrow
'He probably won't sing tomorrow.'
b. I morgen **skal** han antakeligvis ikke synge.
tomorrow shall he presumably NEG sing

The traditional analysis holds that V2 movement requires movement of the finite verb to C^0. Canonical word order in embedded clauses is not V2, as evidenced by the position of the verb with respect to adverbs and negation in (10). This entails that the verb does not move to the embedded C position.

(10) Han er lei for at han antakeligvis ikke **skal** synge i-morgen.
He is sad for that he presumably NEG shall sing tomorrow
'He is sad because he probably won't sing tomorrow.'

It has been observed, however, that V2 word order is possible in some embedded clauses (Julien 2007; Bentzen 2014, a.o.). For example, in (11) the frame adverbial *i morgen* ('tomorrow') has been fronted internal to the embedded clause and the verb has moved past the embedded subject:

(11) Han sa [at i morgen skal han ikke synge.]
He said that tomorrow shall he NEG sing
'He said that TOMORROW he won't sing.'

Sentences like (11) provide evidence for extra functional structure in the left-periphery of the clause under the assumption that *skal* has moved to a head in the CP-domain distinct from the head hosting the complementizer head *at* ('that') and there exists a specifier position between *at* and the verb that *i morgen* can occupy.

In English, embedded fronting of a non-subject does not result in V2/subject-auxiliary inversion, thus observing the absence of V2 in embedded clauses like (12) might provide evidence that the language lacks the extra functional structure.[7]

(12) He said that tomorrow {he will not | *will he not} sing.

Indirect positive evidence might also come from input sentences that do not involve observing different complementizer-level functional structure: English speech errors might also provide relevant evidence of a difference. It is well known that English speakers produce *resumptive* pronouns inside islands to 'rescue' ill-formed sentences (e.g., Ross 1967; Morgan & Wagers 2018). Importantly, English speakers produce resumptives in precisely the locations where Norwegian would allow gaps. For example, the sentences in (11) were observed in natural discourse:

(13) a. There were a bunch of people at the party that I didn't know [who ***they*** were].

b. " . . . the sale of the uranium that nobody knows what **it** means"

–Donald Trump (Gore, Kiely, and Robertson 2016, cited in Morgan & Wagers 2018: 861)

c. "Maybe it was a bad idea to get people together and try to record audio with some equipment *that we didn't know how* ***it*** *worked*."

(CBC Podcast *Personal Best,* Episode: "Know more, Carry Less", ~9:00)

Based on examples such as those in (13), a learner with the knowledge that resumptive pronouns and gaps are in complementary distribution would be able to infer that embedded questions are islands in English. Importantly, drawing inferences based on indirect positive evidence requires non-trivial prior knowledge of the implicational relations between overt forms and (families of) underlying structures.

Indirect positive evidence of the type we describe above is likely to be relatively infrequent in the learner's input. The relative infrequency of such structures may help explain why our participants appear not to have mastered the appropriate generalization and why there is significant inter-individual variation in outcomes despite long-term exposure to and instruction in English. Such effects follow under probabilistic models of grammar competition where conclusively shifting to the subset grammar would require repeated exposure to disconfirmatory evidence (e.g. Yang, 2018).

Conclusion

We have argued that native Norwegian speakers erroneously transfer the grammatical source of *wh-island insensitivity* from their L1 to their L2 English. Such effects are compatible with models of transfer that allow transfer of CP-level functional structure, but not those that restrict transfer to lexical information. We also found evidence that suggested that (some) learners may partially restructure, which we suggested could be triggered by indirect positive evidence. However, our data do not tell us whether the restructuring observed involves transition to the target English analysis or adoption of a divergent compensatory hypothesis that simply ensures closer surface alignment with the English forms.

[7] Of course, the conclusion that English lacks the functional structure is not *forced* by data like (12), which is also compatible with the hypothesis that English has the extra structure, but lacks V2 movement generally. Thus, even these cases do not provide unambiguous indirect evidence. The possibility of vestigial or remnant V2 in English constructions with fronted negative and *only*-phrases further complicate this simple picture.

Acknowledgments
Previous versions of this work were presented at UiT, UMASS, UC Santa Cruz, and at the 2019 CUNY Sentence Processing Conference. We thank audiences for helpful feedback. Special thanks to Jason Rothman for helpful comments on a previous draft. All errors or misrepresentations are our responsibility.

References

Amaral L, and Roeper T (2014) Multiple Grammars and Second Language Representation. *Second Language Research, 30*(1): 3-36.

Anderssen M, Bentzen K, Busterud G, Dahl, A, Lundquist, B, Westergaard, M (2018) The acquisition of word order in L2 Norwegian: The case of subject and object shift. *Nordic Journal of Linguistics.* vol. 41 (3): 247-274.

Ayoun D (1999) Verb movement in French L2 acquisition. Bilingualism: Language & Cognition, 2, 103-125.

Bates D, Mächler M, Bolker B, and Walker S (2014) Fitting linear mixed-effects models using lme4. Fitting Linear. *Journal of Statistical Software*, 67(1):1-48.

Berwick R (1985) *The Acquisition of Syntactic Knowledge*. Cambridge, MA: MIT Press.

Carroll, SE (1995) The irrelevance of verbal feedback to language learning. In L. Selinker, M. Sharwood Smith, W. E. Rutherford, & L. Eubank (Eds.), *The Current State of Interlanguage : Studies in Honor of William E. Rutherford*. Amsterdam: John Benjamins Publishing Co.

Carroll, SE (2001) *Input and Evidence: The Raw Material of Second Language Acquisition,* Amsterdam: John Benjamins.

Chomsky N (1977) On wh-movement. In Culicover P, A Akmajian, and T Wasow (Eds.), *Formal syntax*, 71-132. New York: Academic Press.

Chomsky N and Lasnik H (1977) Filters and control. *Linguistic Inquiry*, *8*(3): 425-504.

Clahsen H and Muysken P (1986) The availability of universal grammar to adult and child learners: a study of the acquisition of German word order. *Second Language* 2(2): 93-119.doi: 10.1177/026765838600200201

Dahl A (2004) Negative evidence in L2 acquisition. *Nordlyd* 32(1): 28-45. doi: doi.org/10.7557/12.57

Eubank L (1994). Optionality and the Initial State in L2 Development. In: T Hoekstra and B Schwartz (eds.) *Language Acquisition Studies in Generative Grammar*. Amsterdam: John Benjamins, pp. 369-388.

Felser C, Cunnings I, Batterham C, and Clahsen H (2012) The timing of island effects in nonnative sentence processing. *Studies in Second Language Acquisition*, *34*(1): 67-98.

Foraker S, Regier T, Khetarpal N, Perfors A, and Tenenbaum J (2009). Indirect evidence and the poverty of the stimulus: The case of anaphoric "one." *Cognitive Science*, 33: 287-300. doi: 10.1111/j.1551-6709.2009.01014.x

Gore D, Kiely E, and Robertson L (2016) Spinning the FBI letter. FactCheck.org. Accessed on 25 February 2017. http://www.factcheck.org/2016/10/spinning-the-fbi-letter/.

Johnson J and Newport E (1991) Critical period effects on universal properties of language: The status of subjacency in the acquisition of a second language. *Cognition* 39(3): 215-258. doi: 10.1016/0010-0277(91)90054-8

Judy, T, and Rothman, J (2010). From a superset to a subset grammar and the Semantic Compensation Hypothesis: Subject pronoun and anaphora resolution evidence in L2 English. In K. Franich, K. M. Iserman, & L. L. Keil (Eds.), BUCLD 34: Proceedings of the 34th Boston University conference on language development (pp. 197–208). Somerville, MA: Cascadilla Press.

Kim E, Baek S and Tremblay A (2015) The Role of Island Constraints in Second Language Sentence Processing. *Language Acquisition*, 22(4), 384-416. doi:10.1080/10489223.2015.1028630

Kush D, Lohndal T, and Sprouse J (2018) Investigating variation in island effects: A case study of Norwegian wh-extraction. *Natural Language & Linguistic Theory* 36: 743-779

Kush D, Lohndal T, and Sprouse J (2019) On the island sensitivity of topicalization in Norwegian: An experimental investigation. *Language* 95: 393-420.

Kuznetsova A, Brockhoff PB, Christensen RHB (2017) lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, 82(13) 1–26. doi: 10.18637/jss.v082.i13

Li, X (1998) Adult L2 Accessibility to UG: An issue revisited. In: Flynn S, Martohardjono G and O'Neil W (eds) *The Generative Study of Second Language Acquisition.* Lawrence Erlbaum Associates : Mahwah, NJ.

Lindahl, F. 2017. Extraction from relative clauses in Swedish. Doctoral dissertation, University of Gothenburg: Gothenburg, Sweden.

Maling J and Zaenen A (1982) A phrase structure account of Scandinavian extraction phenomena. In Jacobson P & G Pullum (eds.), *The nature of syntactic representation,* 229-282. Dordrecht: Springer.

Martohardjono G (1993) Wh-movement in the acquisition of a second language: A cross-linguistic study of three languages with and without movement. Doctoral dissertation: Cornell University.

Mazurkewich I (1984) The acquisition of the dative alternation by second language learners and linguistic theory. *Language Learning* 34(1): 91-108. doi: 10.1111/j.1467-1770.1984.tb00997.x

Omaki A and Schulz B (2012) Filler-gap dependencies and island constraints in second language sentence processing. Studies in Second Language Acquisition, 33(4): 563-588

Pearl LS. and Mis B (2016) The role of indirect positive evidence in syntactic acquisition: A look at anaphoric one. *Language*, *92*(1): 1-30.

Perez-Leroux AT and Li X (1998) Selectivity in the acquisition of complex NP islands. In: E Klein and G Martohardjono (eds)*The Development of Second Language Grammars: A Generative approach*. Amsterdam: John Benjamins, pp. 148–68.

Perfors A, Tenenbaum JB, and Regier T (2011) The learnability of abstract syntactic principles. *Cognition*, *118*(3): 306-338.

Perlmutter DM (1971) *Deep and surface structure constraints in syntax*. New York: Holt, Rinehart and Winston.

Plough I (1992) Indirect Negative Evidence, Inductive Inferencing and Second Language Acquisition. In: L Eubank, L Selinker and M Sharwood Smith (eds) *The Current State of Interlanguage: Studies in Honor of William E. Rutherford*, Amsterdam: John Benjamins, pp. 89-105.

Ramscar M, Dye M, and McCauley SM (2013) Error and expectation in language learning: The curious absence of" mouses" in adult speech. *Language*, 760-793.

Rankin T (2012) The transfer of V2: Inversion and negation in German and Dutch learners of English. *International Journal of Bilingualism*, 16(1), 139-158. doi:10.1177/1367006911405578

Rankin T (2014) Variational learning in L2: The transfer of L1 syntax and parsing strategies in the interpretation of wh-questions by L1 German learners of L2 English. *Linguistic Approaches to Bilingualism* 4(4): 432-461 doi:10.1075/lab.4.4.02ran

Regier T and Gahl S (2004) Learning the unlearnable: The role of missing evidence. *Cognition*, *93*(2): 147-155.

Reinhart T (1981) A second COMP position. *Theory of Markedness in Generative Grammar*, ed. by Adriana Belletti, 518-557. Pisa: Scuola Normale Superiore.

Rizzi L (1982) *Issues in Italian Syntax*. Dordrecht: Foris.

Rohde DL and Plaut DC (1999) Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, *72*(1): 67-109.

Ross JR (1967) Constraints on variables in syntax. Doctoral dissertation, MIT, Cambridge, MA.

Rothman, J., Alonso, J. G., & Puig-Mayenco, E. (2019). *Third language acquisition and linguistic transfer* (Vol. 163). Cambridge University Press.

Rothman J & Iverson M (2013). Islands and objects in L2 Spanish: Do You Know the Learners Who Drop?. *Studies in Second Language Acquisition*, *35*(4), 589-618.

Samordna opptak (2019) *Krav til norsk og engelsk*. https://www.samordnaopptak.no/info/utenlandsk_utdanning/sprakkrav/krav-til-norsk-og-engelsk-for_hoyere_utdannning/index.html

Schwartz B and Sprouse R (1994) Word order and nominative case in nonnative language acquisition: a longitudinal study of (L1 Turkish) German interlanguage. In: T Hoekstra and B Schwartz (eds) *Language Acquisition Studies in Generative Grammar*. Amsterdam: John Benjamins, pp. 317-368.

Schwartz B and Sprouse R (1996) L2 cognitive states and the Full Transfer/Full Access model. *Second Language Research*, 12(1): 40–72.

Snyder W (2007) *Child Language: The Parametric Approach* Oxford: Oxford University Press.

Sorace A (1996) The use of acceptability judgments in second language acquisition research. In: Ritchie WC and Bhatia TK (eds) *Handbook of Second Language Acquisition*. San Diego, CA: Academic Press, pp. 375–409.

Spinner P and Gass SM (2019) *Using Judgments in Second Language Acquisition Research:* New York: Routledge.

Sprouse J (2007) A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge. Doctoral dissertation, University of Maryland.

Sprouse J, Fukuda S, Ono H, and Kluender R (2011) Reverse island effects and the backward search for a licensor in multiple *wh*-questions. *Syntax* 14:179-203.

Sprouse J, Wagers M, and Phillips C. (2012) A test of the relation between working memory and syntactic island effects. *Language* 88.82-124.

Sprouse J, Caponigro I, Greco C, and Cecchetto C (2016) Experimental syntax and the variation of island effects in English and Italian. *Natural Language and* Linguistic Theory 34:307-344.

Trahey, M., & White, L. (1993). Positive evidence and preemption in the second language classroom. *Studies in second language acquisition*, *15*(2), 181-204.

Trapman M and Kager R (2009) The acquisition of subset and superset phonotactic knowledge in a second language. *Language Acquisition* 16(3): 178–221. doi.org/10.1080/10489220903011636

Vainikka A and Young-Scholten M (1994) Direct access to X'-theory: Evidence from Korean and Turkish adults learning German. In: T Hoekstra and B Schwartz (eds) *Language Acquisition Studies in Generative Grammar*. Amsterdam: John Benjamins, pp. 265-316.

Vainikka A and Young-Scholten M (1996) Gradual development of L2 phrase structure. *Second Language Research* 12(1) 7-39. doi: 10.1177/026765839601200102

Vikner S, Christensen KR, and Nyvad, AM (2017) V2 and cP/CP: *Order and structure in syntax I: Word order and syntactic structure*, ed. by Laura R. Bailey and Michelle Sheehan, 313-324. Berlin: Language Science Press.

Westergaard M (2003) Unlearning V2: Transfer, markedness, and the importance of input cues in the acquisition of word order in English by Norwegian children. *EUROSLA Yearbook* 3: 77-101.

Westergaard, M (2014). Linguistic variation and micro-cues in first language acquisition. *Linguistic Variation* 14(1). 10.1075/lv.14.1.02wes.

Westergaard, M (2019). Microvariation in multilingual situations: The importance of property-by-property acquisition. *Second Language Research*. https://doi.org/10.1177/0267658319884116

White L (1989). Linguistic universals, markedness and learnability: Comparing two different approaches. *Second Language Research* 5(2): 127-140. doi: 10.1177/026765838900500202

Wexler K and Manzini MR (1987) Parameters and learnability in binding theory. In: Roeper T and Williams E (eds) *Parameter Setting*. Dordrecht: D. Reidel, pp. 41-76.

White L (1989a) *Universal Grammar and second language acquisition.* Amsterdam: John Benjamins.

White L (1989b) The principle of adjacency in second language acquisition: do L2 learners observe the subset principle? In: S Gass and J Schachter (eds) Linguistic Perspectives on Second Language Acquisition. Cambridge: Cambridge University Press, pp- 134-158.

White L (2003) *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.

White L and Genesee F (1996) How native is near-native? The issue of age and ultimate attainment in the acquisition of a second language. *Second Language Research* 12(3): 233-265. doi: 10.1177/026765839601200301

White, L., & Juffs, A. (1998). Constraints on Wh- movement in two different contexts of non-native language acquisition: Competence and processing. In: S Flynn, G Martohardjono and W O'Neill (eds) *The generative study of second language acquisition* Hillsdale, NJ: Lawrence Erlbaum Associates, pp. 111–130.

Wolfe Quintero K (1992) Learnability and the acquisition of extraction in relative clauses and wh-questions. *Studies in Second Language Acquisition* 14(1): 39–70. doi:10.1017/S0272263100010469

Yang C (2018) A formalist perspective on language acquisition. *Linguistic Approaches to Bilingualism* 8(6): 665-706. doi:10.1075/lab.18014.yan

Yuan B (1997) Asymmetry of null subjects and null objects in Chinese speakers L2 English. *Studies in Second Language Acquisition* 19(4): 467–497. doi: 10.1017/S0272263197004038

Zobl H (1988) Configurationality and the subset principle: the acquisition of V' by Japanese learners of English. In: Pankhurst J, Sharwood Smith M and Van Buren P (eds) *Learnability and Second languages: a Book of Readings.* Dordrecht: Foris, pp. 116-131.

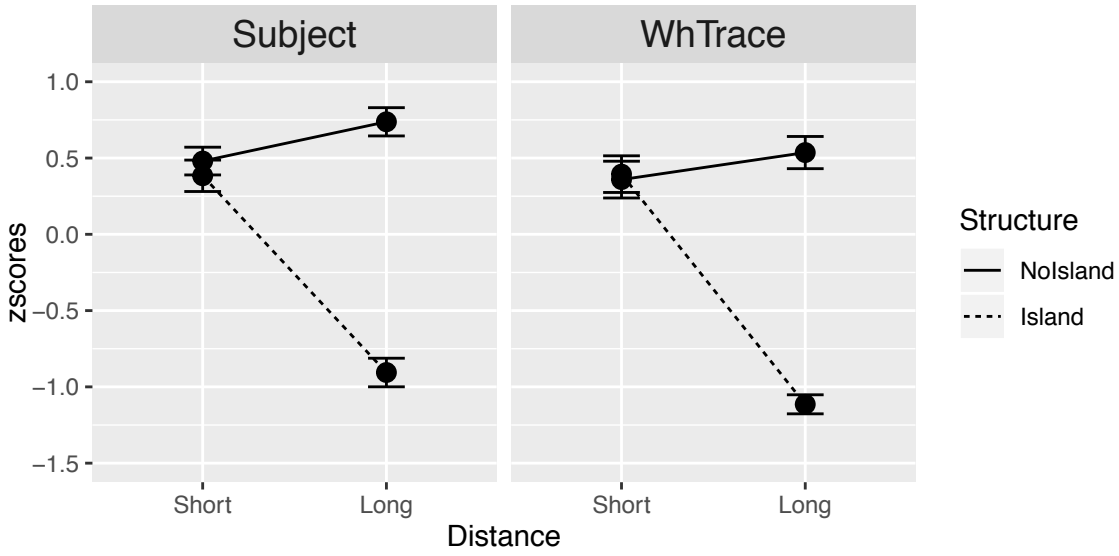


Figure 1. Average z-scored acceptability judgments from native English control participants in the Subject island (left panel) and Wh-Trace island (right panel) sub-experiments.

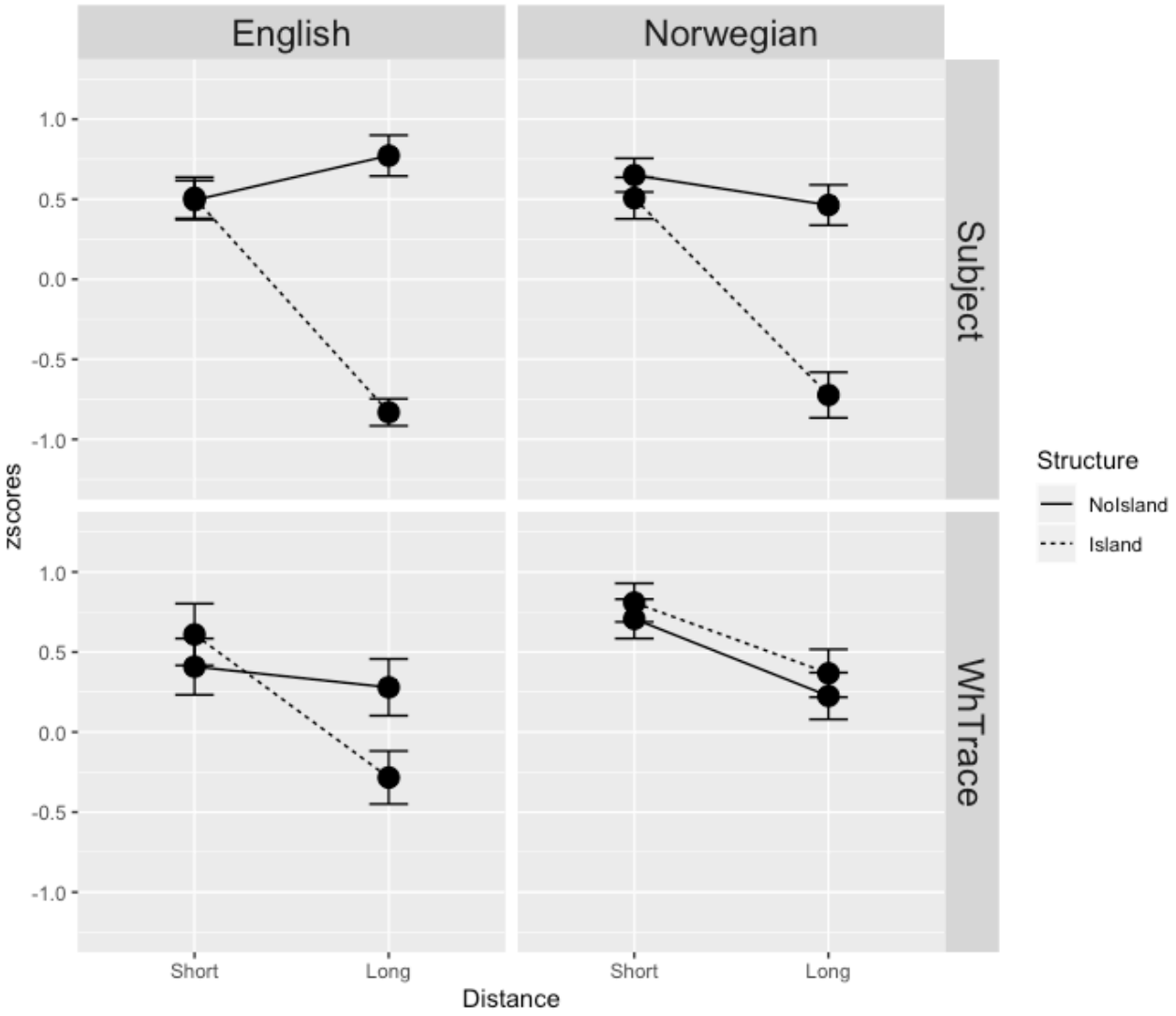


Figure 2. Average z-scored acceptability judgments from Norwegian participants in Experiment 1. Rows correspond to the island items judged and columns correspond to the language of presentation.

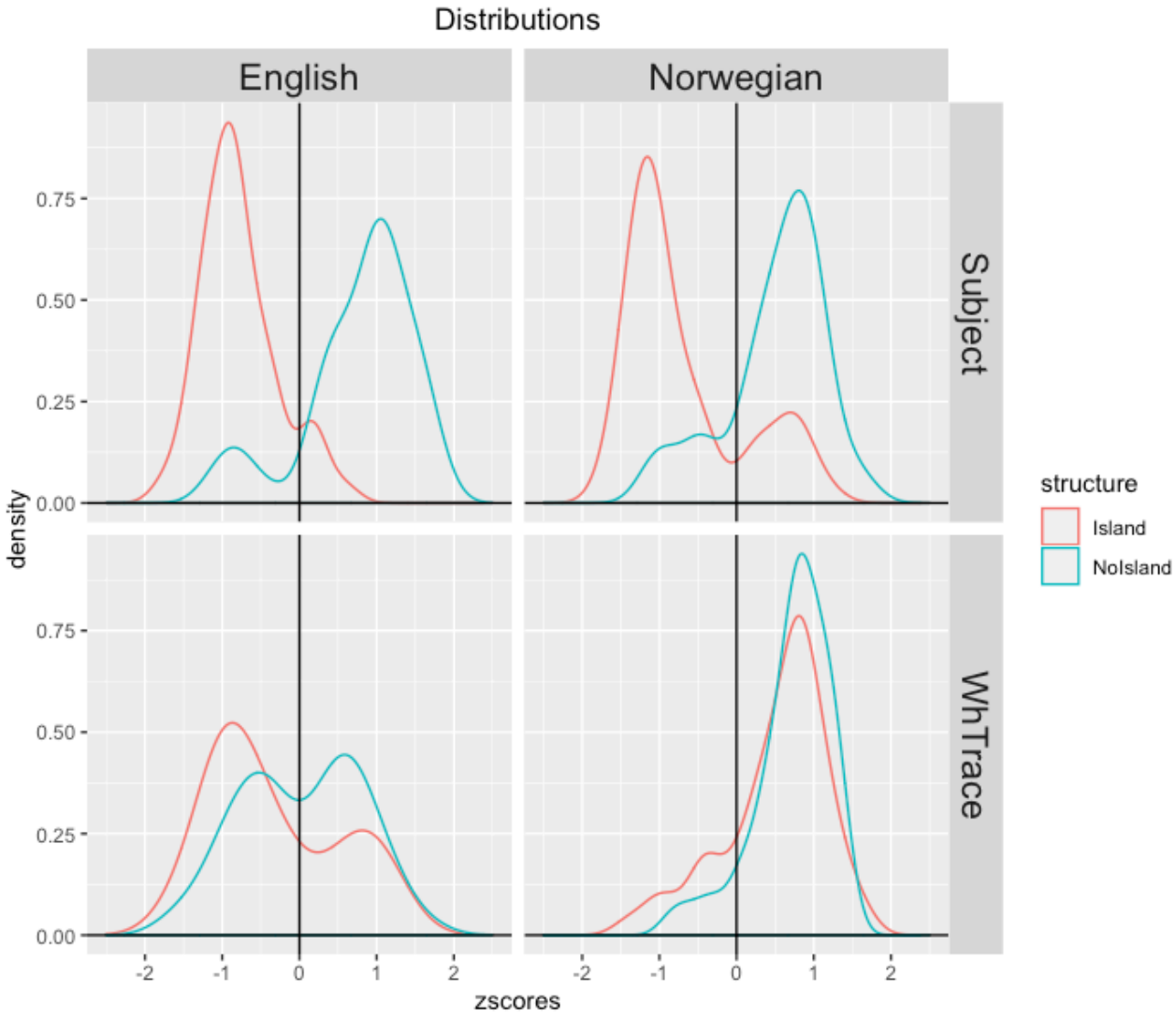


Figure 3. Distribution of judgments in Long-NoIsland and Long-Island conditions for each island and language pair in Experiment 1b.

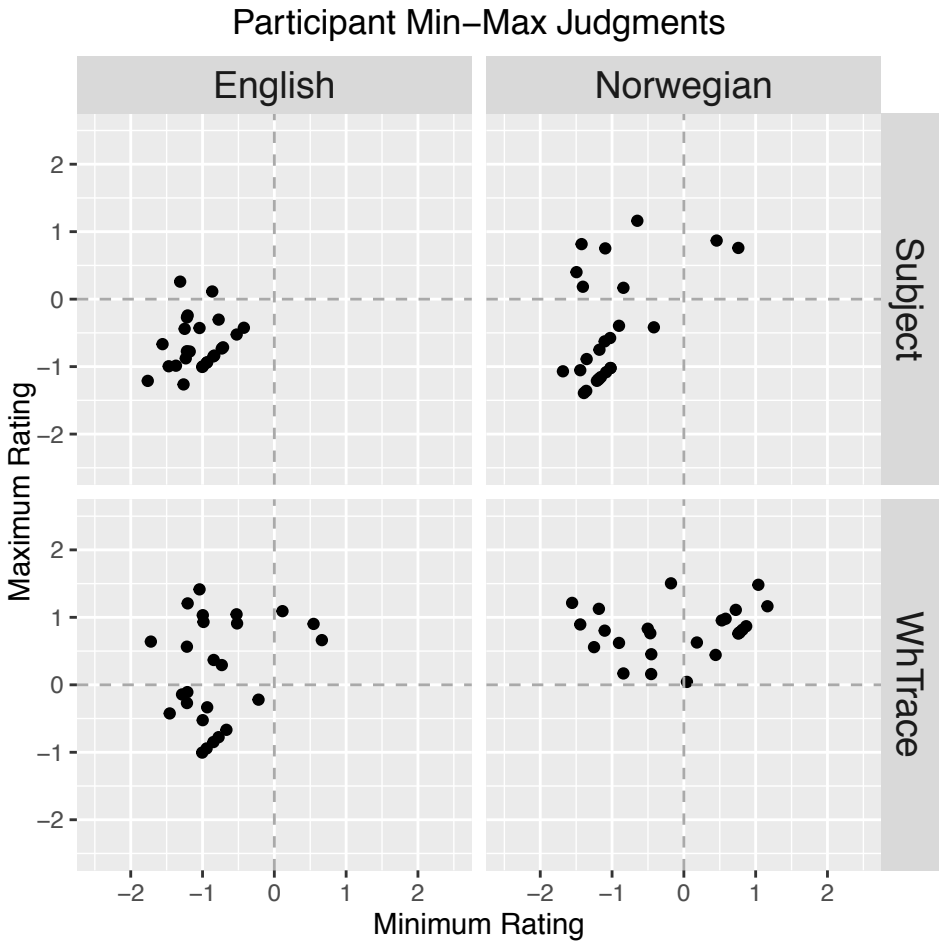


Figure 4. Plots of by-participant minimum and maximum judgments for each island-language pair in Experiment 1b. Each dot represents a single participant.

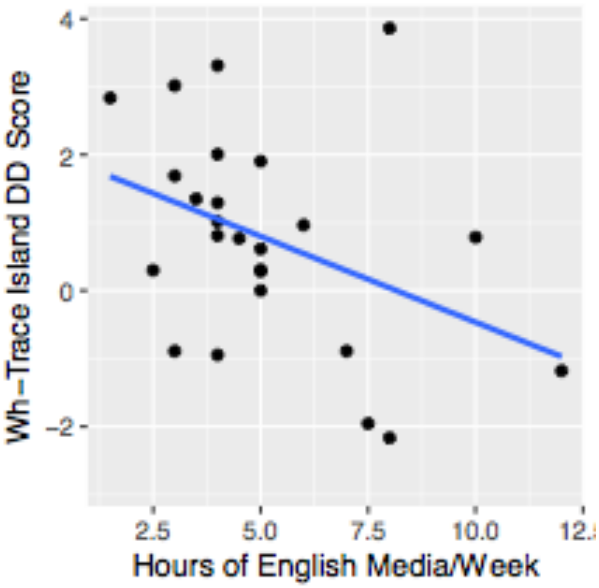


Figure 5. Correlation between participant Wh-Trace DD scores and self-reported hours of English media exposure in Experiment 1b.

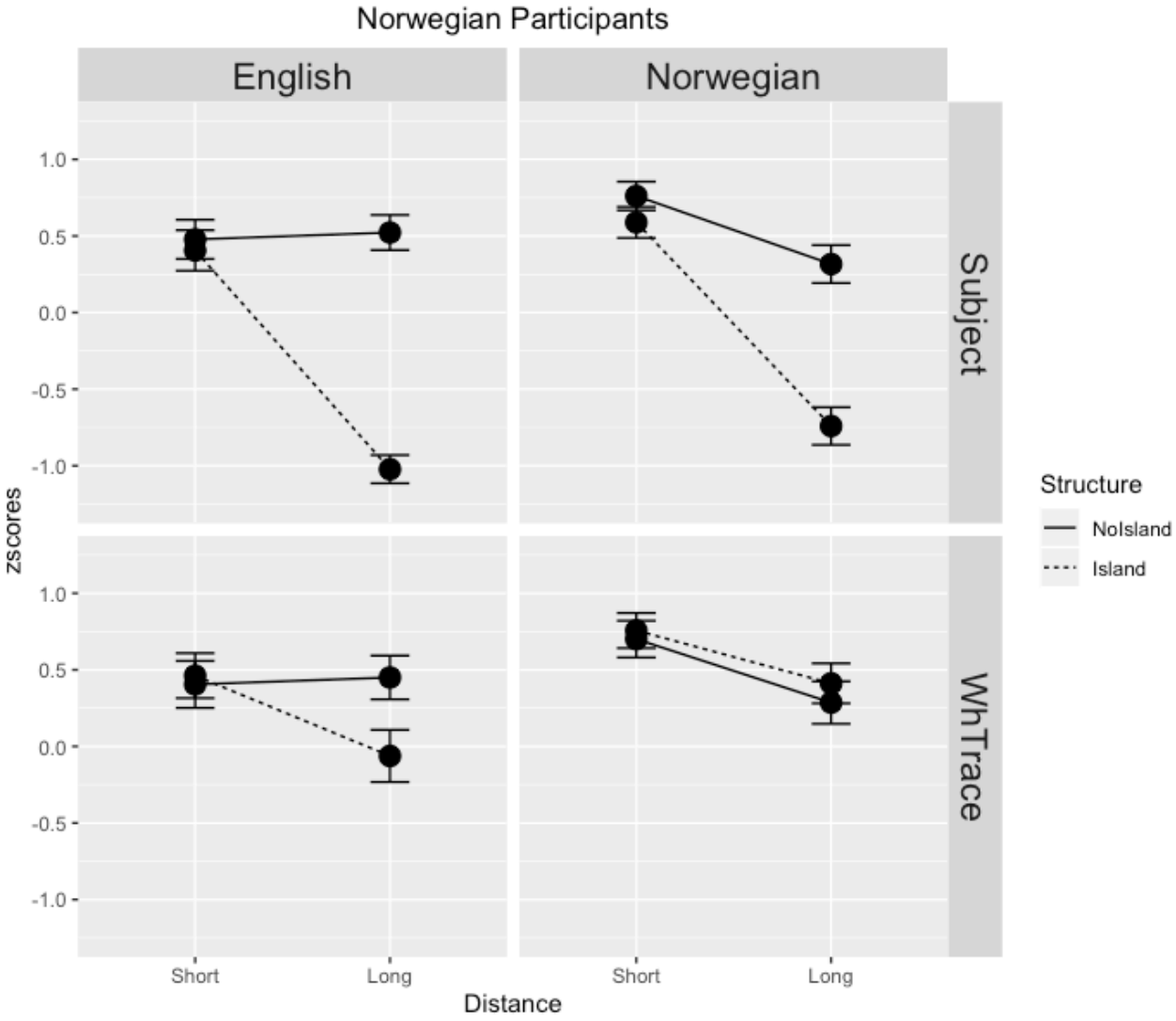


Figure 6. Average z-scored acceptability judgments from Norwegian participants in Experiment 2. Rows correspond to the island judged and columns correspond to the language of presentation.

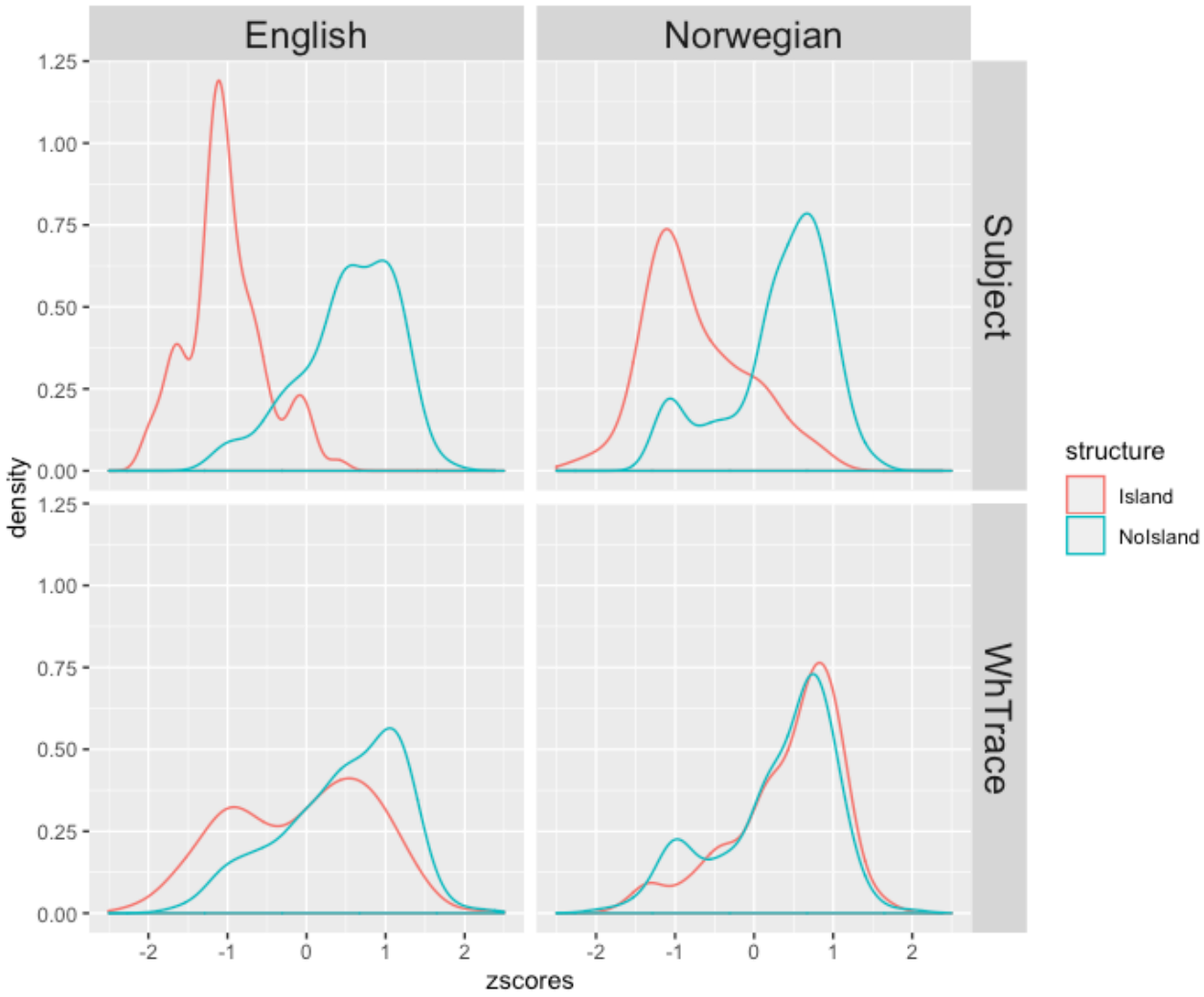


Figure 7. Distribution of judgments in Long-NoIsland and Long-Island conditions for each island and language pair in Experiment 2.

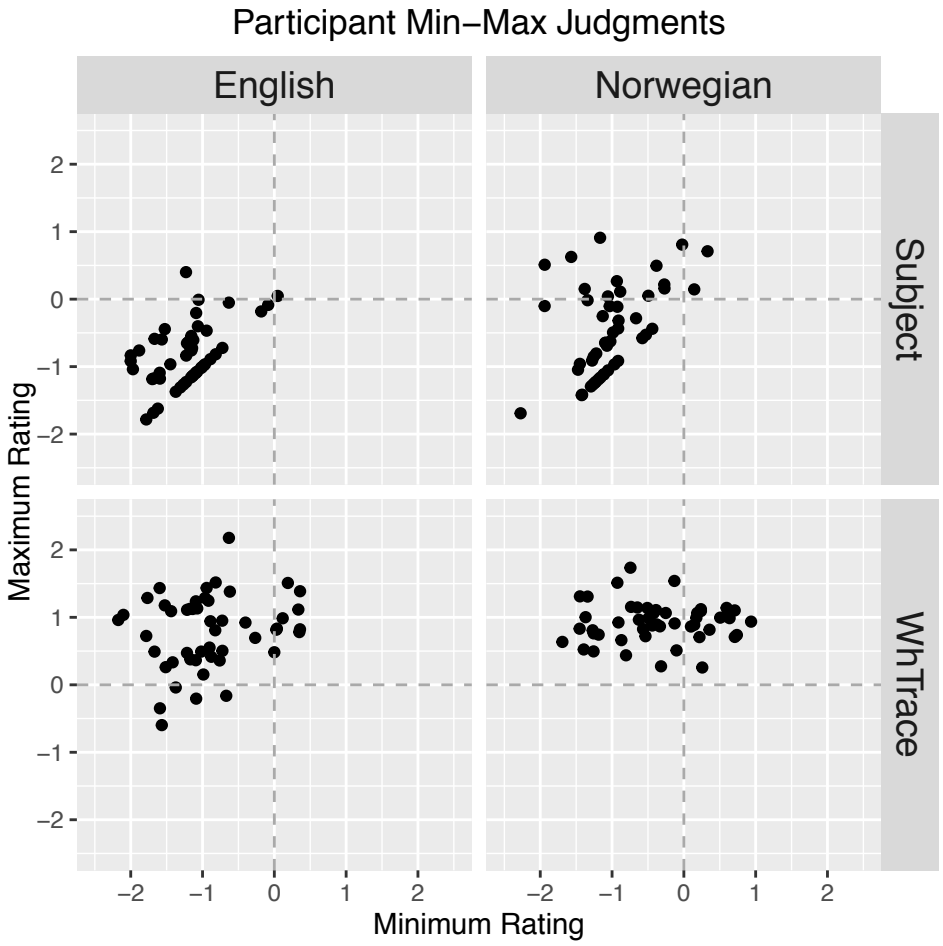
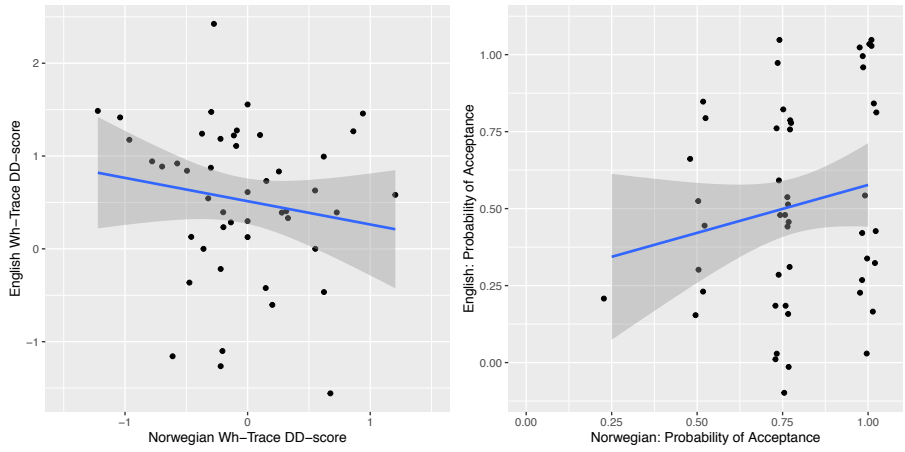


Figure 8. Plots of by-participant minimum and maximum judgments for each island-language pair in Experiment 2. Each dot represents a single participant.



(A) (B)

Figure 9. (A) Correlation between individual participants' Wh-Trace DD scores in Norwegian and English in Experiment 2. Each dot represents a single participant. (B) Correlation between individual participants' probability of accepting a Wh-Trace island violation in Norwegian and English. Dots are jittered.