The That-Trace Effect and Island Boundary-Gap Effect are the same: Demonstrating equivalence with Null Hypothesis Significance Testing and psychometrics

Adam M. Morgan
NYU School of Medicine
223 E 34th St, New York NY 10016
adam.morgan@nyulangone.org

Abstract This paper demonstrates a novel approach in experimental syntax, leveraging *psychometric* methods to resolve a decades-old puzzle. Specifically, gaps in subject position are more acceptable than gaps in object position in non-islands, while the reverse is true in islands (the *Island Boundary-Gap Effect*). Attempts at explaining this asymmetry generally attribute it to a violation of the same constraint that renders gaps unacceptable after the overt complementizer '*that*' (the *That-Trace Effect*). However, the two effects involve distinct syntactic structures, and there is no a priori reason to believe they are the same beyond the elegance of a unified account. One limitation has been the difficulty of testing for equivalence in the Null Hypothesis Significance Testing framework: when two constructs behave similarly, it generally constitutes an uninterpretable null result. Experiments 1 and 2 use standard experimental methods to circumvent this problem, but ultimately provide evidence that is at best just consistent with equivalence. Experiment 3 demonstrates a novel approach which shows that individual differences in the That-Trace Effect correlate with individual differences in the Island Boundary-Gap Effect, after removing correlated variance from carefully-chosen controls. This psychometric approach provides positive evidence that the two effects do indeed derive from the same underlying phenomenon.

Keywords: syntax; That-Trace Effect; individual differences; psychometrics; Null Hypothesis Significance Testing

1 Introduction

Linguistic models generally aim to reduce the number of unexplained phenomena that must be independently accounted for. One example of this involves two syntactic phenomena: the *That-Trace Effect* and what will be referred to here as the *Island Boundary-Gap Effect*. On the surface, these effects appear remarkably similar. A central assumption of nearly all models of syntax for the last half-century has therefore been that they result from the same underlying constraint (e.g. Chomsky 1981; McDaniel et al. 2015; Chomsky 2014).

However, the That-Trace Effect and Island Boundary-Gap Effect in fact involve distinct structures and lexical categories, making it far from clear that a unified analysis is warranted. This paper provides the first evidence (to my knowledge), aside from surface-level similarity, that the two effects derive from the same underlying constraint.

The That-Trace Effect and the Island Boundary-Gap Effect are both properties of *long-distance syntactic dependencies*. In such dependencies, as in the relative clauses in (1), a noun appears far away from its canonical position:

(1) a. *Subject gap*
…the patient [that I think [__ took the pill]]
b. *Object gap*
…the pill [that I think [the patient took __]]

That is, whereas in ordinary sentences like *The patient took the pill*, subjects closely precede the verb and objects closely follow it, in (1a), '*patient*' is the subject of '*took*,' but the two words are separated by two clause boundaries (brackets). Similarly, in (1b), '*pill*' is the object of '*took*,' but these are also separated by two clause boundaries. In both cases, the canonical position for the noun is left empty. These empty positions are referred to as *gaps* (or *traces*; represented with underscores throughout). Example (1a) demonstrates a *subject gap*, or a gap in subject position, and example (1b) demonstrates an *object gap*.

The That-Trace Effect is exemplified by the unacceptability of (1a) when the complementizer '*that*' appears before the gap, as in (2) (Perlmutter 1968):

(2) *That-Trace Effect*
*…the patient [that I think [**that** __ took the pill]]

While '*that*' is typically optional in English (e.g., *I think (that) the patient took the pill*), in (2) its presence renders the structure ungrammatical.

There are reasons to think the That-Trace Effect may reflect something deep about human language. First, despite initially seeming to be an idiosyncratic quirk of English grammar, it has in fact turned out to be cross-linguistically common, documented in a number of languages across several families including French (Perlmutter 1968), Russian (Pesetsky 1982), Levantine Arabic (Kenstowicz 1989), German (Featherston 2005), Nupe (Kandybowicz 2006), and Wolof (Martinović 2013).

Second, the distribution of the That-Trace Effect in speech poses a problem for language acquisition. Speakers almost never produce '*that*' in sentences like (1a) (i.e., subject relative clauses), where it would be ungrammatical. But, somewhat surprisingly, speakers also almost never produce '*that*' in sentences like (1b) (object relative clauses), even though in these cases it is in fact grammatical (Phillips 2013). Despite similar rates of '*that*' production in their input, children must learn that '*that*' is optional in the latter, but unacceptable in the former. This hints at the possibility that more information than is contained in the input may be brought to bear on the acquisition of the That-Trace Effect.

To understand the Island Boundary-Gap Effect, a bit of background on gap processing is relevant. Extensive literature indicates that, relative to subject gaps, object gaps are harder to process (Ford 1983; Hakes et al. 1976; Holmes & O'Regan 1981), less frequent (McDaniel et al. 1998), cross-linguistically less common (Keenan & Comrie 1977), and are sometimes reported to have lower acceptability (Han et al. 2012; Morgan & Wagers 2018). However, the pattern of acceptability reverses in structures known as *islands*.

Islands, exemplified in (3), are defined by the fact that they are unacceptable when gaps appear inside them (Ross 1967). Unlike in non-islands like (1), in islands subject gaps are less acceptable than object gaps (Chomsky 1981; Pesetsky 1982). This decreased acceptability of gaps when they are in subject position in an island – that is, when they immediately follow an island boundary – is the Island Boundary-Gap Effect.

(3) a. *Subject gap in island*
**…the patient [that I wonder [when __ took the pill]]
b. *Object gap in island*
*…the pill [that I wonder [why the patient took __]]

1.1 A unified account

One way to account for the puzzling asymmetry between subject and object gaps in islands and non-islands might be to stipulate a specific constraint against subject gaps in islands. However, researchers have long noted that the That-Trace Effect and the Island Boundary-Gap Effect are suspiciously similar (Chomsky 1980), at least on the surface. They both involve an unacceptable subject gap immediately following an overt clausal function word, like '*that*' (2) or '*why*' (3a). Drawing on this surface similarity, many attempts have been made to attribute the That-Trace Effect and the Island Boundary-Gap Effect to the same underlying constraint.

For instance, McDaniel et al. (2015) offer an account based on the idea that processing pressures during production eventually become grammaticized over the course of language evolution (Hawkins 2014). They stipulate that clause-initial function words like *that* and *whether* demarcate separate planning units for the language production system. They then argue that connecting clauses with a long-distance dependency is more difficult immediately after the beginning of a new planning unit. Over the course of language evolution, this difficulty results in the grammaticization of a constraint against gaps immediately after a clause-initial function word.

Perhaps the most famous attempt at unification is a linguistic constraint proposed by Chomsky (1981) known as the Empty Category Principle (ECP). According to the ECP, which is understood to be a general principle that applies in all languages, gaps are prohibited from appearing immediately after clause-initial function words like *that*, *when*, *because*, etc. The specifics of this proposal require extensive knowledge of Government and Binding (Chomsky 1981), a prominent theoretical framework from the 1980s, but are not critical for the present purposes (see Chomsky 1981; Lasnik & Saito 1984; Kayne 1981 for further detail). Like McDaniel et al.'s account, the ECP also explains the ungrammaticality of the That-Trace Effect and the Island Boundary-Gap Effect as the result of violating a single underlying constraint.

In spite of the elegance of a unified explanation of the That-Trace Effect and the Island Boundary-Gap Effect, there is no a priori reason to believe that the two should be the same. Indeed, there are several key differences between the two. While the That-Trace Effect is conditioned on the presence of a specific complementizer, '*that*,' the Island Boundary-Gap Effect is conditioned on any type of island boundary, which may or may not involve a complementizer. Island boundaries *can* be marked with a complementizer (or complementizer-like function word; the particular analysis does not matter here) like *if* or *because* (4a). But they can also be marked with a relative pronoun or *wh*-operator like *who*, *why* or *when* (4b), or even a *wh*-operator involving an entire referential phrase like *which of these staplers* (4c). Not only are these latter examples not complementizers, they also occupy a different syntactic position within the sentence,[1] as shown in Table 1.

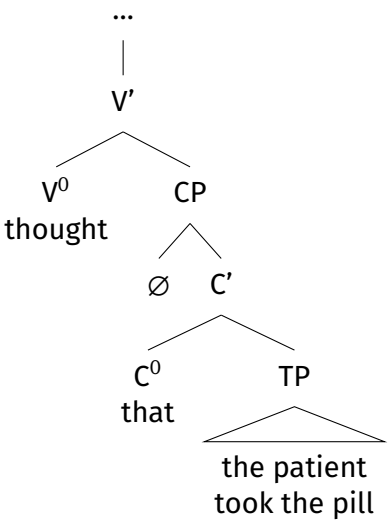
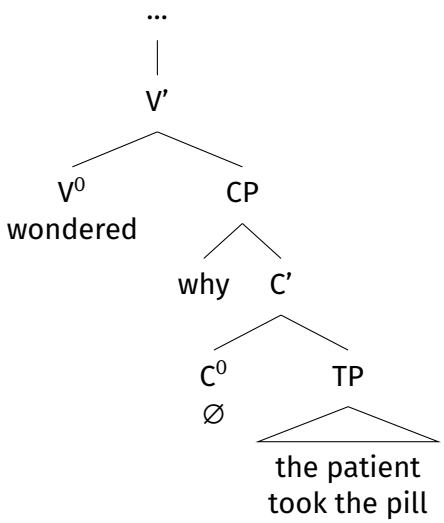
(4) Distinct types of island boundaries

a. *Complementizer*
I wonder [*if* the patient took the pill].

[1] Specifically, while complementizers are the head (C^0) of a complementizer phrase (CP), relative pronouns and wh- operators like *who*, *which* and *why* appear in the specifier position of CP, which is structurally above C^0. The existence of these two positions is made especially clear by dialects of English like that spoken in Belfast, which allows relative pronouns and complementizers to co-occur, as in "They didn't know which model that they discussed" (Baltin 2010: 331, Ex. 1; see also van Gelderen 2013 and Rizzi & Shlonsky 2008: 131, Ex. 30 for Québec French).

b. *Wh- operator*
I wonder [*why* the patient took the pill].
c. *Wh- operator with internal structure*
I wonder [*which of these staplers* the patient threw at the nurse].

Table 1: Illustrations of the distinct positions that complementizers and relative pronouns occupy.

Complementizer (*that*)	*Wh-* operator (*why*)
	
"...thought that the patient took the pill"	*"...wondered why the patient took the pill"*

Thus, the similarity between the That-Trace Effect and the Island Boundary-Gap Effect is superficial, at best. This relatively weak evidence has nonetheless been used to motivate the assumption, which is pervasive throughout the literature, that the two effects are the same. This assumption is clearly appealing: it reduces the number of stipulative constraints in the grammar. But it only does so by one. Without stronger evidence in favor of unification, it is not clear that such a small reduction is sufficient to justify such a consequential assumption.

The present aim is to experimentally assess whether the two effects are the same, and in so doing, demonstrate a novel approach to questions about equivalence. As will be seen in Experiments 1 and 2, the type of experimental paradigms and statistics standardly used in experimental syntax impose a familiar constraint: it is easy to design a study intended to detect *differences* between constructs, but showing that two constructs are the same is trickier. To do so, Experiment 3 leverages psychometric methods to flip the usual logic of the experimental design, so that a significant result can provide evidence of equivalence rather than difference.

1.2 The joint logic of experimental methods and Null Hypothesis Significance Testing

The standard experimental approach in the cognitive sciences is designed to reveal differences between conditions. *Equivalence* between conditions, even though it is often useful to know about, is harder to demonstrate. This difficulty derives from the fact that *Null*

Hypothesis Significance Testing, the dominant statistical framework, is tailored to look for differences between measures, not similarities.[2]

But it is not *necessarily* the case that Null Hypothesis Significance Testing cannot be used to test for equivalence. To foreshadow, Experiment 3 leverages the insight from economics and neuroscience that any systematic *microvariation* in an underlying phenomenon should be detectable in the surface phenomena it causes (Wiener 1956; Granger 1969). Specifically, if the That-Trace Effect and the Island Boundary-Gap Effect do in fact derive from the same underlying constraint, and if there are individual differences (i.e., microvariation) in that underlying constraint, then those individual differences should be present in both effects. Experiment 3 compares individual differences in the acceptability of the two effects to determine whether they correlate beyond what can be explained by extraneous sources of individual differences in acceptability judgment data.

This logic relies on a distinction between statistical hypotheses, which are about variables – the measured values – and theoretical hypotheses, which are about psychological constructs – the black-box entities that variables are often meant to stand in for. The statistical null hypothesis in Null Hypothesis Significance Testing is always that two variables do not differ, and the corresponding theoretical hypothesis is *usually* that two constructs do not differ. But the hypotheses do not have to align in this way.

This is easy to overlook because variables and constructs are so often confounded. In the standard experimental approach, dependent variables tend to be measures of cognitive constructs like grammaticality, working memory capacity, or attention span. When interpreting the data, these variables are treated as proxies for the corresponding construct, and the line between the two is often neither clear nor consequential.

By analyzing individual differences in constructs rather than the constructs themselves, Experiment 3 formulates a statistical question such that a difference between variables results when two psychological constructs behave in the same way. In such cases, a positive statistical result can constitute evidence for equivalence.

The usual difficulty in assessing equivalence is therefore not an inherent limitation of Null Hypothesis Significance Testing, but a limitation of the joint logic of Null Hypothesis Significance Testing and the standard approach to experimental design. Experiment 3 circumvents this issue by employing a *psychometric* experimental design.

Psychometrics is the subfield of psychology that deals with how to measure psychological constructs. Rather than being concerned with the nature of these constructs (e.g., whether language processing uses domain-general or domain-specific mechanisms), its aim is to measure a construct as accurately as possible (Cronbach & Meehl 1955). Psychometric experiments rarely contain experimental manipulations, and do not often attempt to measure within-subjects changes. Instead, the goal is to measure psychological states as they already are (Swets et al. 2007; Rust & Golombok 2014; Furr 2017). As a field, its primary occupation has been with the development of testing, as in college entrance exams, IQ tests, professional licensing, workplace aptitude, etc. (Crocker & Algina 1986).

Psychometric methods often differ considerably from standard methods, doing away with practices like counterbalancing and randomization to avoid introducing unwanted

[2] There do exist frameworks for assessing the null hypothesis, such as Bayes Factor Analysis (Kass & Raftery 1995) (although it is hard to imagine how one might apply Bayes Factor here due to the many extraneous differences between the structures). There are also some existing approaches to demonstrating equivalence with Null Hypothesis Significance Testing. In pharmacological research, the Two One-Sided t-Test is commonly used to provide evidence for equivalence. However, a more accurate characterization of this test is that it provides evidence that two conditions are similar enough so as to be inconsequential, but not necessarily that they are equivalent. See also Hoenig & Heisey (2001) on arguing for the null when power is sufficiently high.

sources of variability. Outside of work on individual differences, psychometrics has had dwindling contact with cognitive science in recent years, although it has on occasion been adopted to divide experimental subjects into groups based on different cognitive abilities (Swets et al. 2007; Longo et al. 2008). The use of psychometric methods in the present study to demonstrate equivalence is, to my knowledge, a novel contribution.

1.3 The present experiments

Here, three experiments with large sample sizes aim to provide further evidence for unifying the two effects. Experiment 1 ($N = 161$) examines acceptability while deconfounding the subject position in *wh*-islands from the position immediately after clause-initial function words. Experiment 2 ($N = 189$) aims to validate that the types of stimuli from Experiment 1, which have not been previously used as stimuli in experimental research, show prototypical properties of islands. While the findings of Experiments 1 and 2 are *consistent* with a unified theory of the That-Trace Effect and the Island Boundary-Gap Effect, the experiments are unable to definitively demonstrate that the effects are the same due to inherent limitations on the joint logic of the standard experimental design and Null Hypothesis Significance Testing.

Experiment 3 ($N = 104$) circumvents these limitations using a novel approach in experimental syntax. Leveraging psychometric methods, it directly tests for equivalence by analyzing individual differences in the That-Trace Effect and the Island Boundary-Gap Effect to determine whether the two correlate beyond what would be expected due to extraneous shared properties.

2 Experiment 1

One obstacle to determining whether the That-Trace Effect and the Island Boundary-Gap Effect reflect the same underlying constraint is that it is difficult to dissociate potential causes of the Island Boundary-Gap Effect. That is, it might be that there is a specific constraint against subject gaps in islands. But it might also be that there is a general constraint against subject gaps after clause-initial function words like *that* and *when*, as proposed by Chomsky (1981) and McDaniel et al. (2015). The difficulty in distinguishing between these possibilities is due to the fact that clause-initial function words at island boundaries are obligatory (e.g., the '*why*' in "…who I wondered [why he attacked __]"). Thus, the subject position in islands is confounded with the position immediately following a clause-initial function word.

Experiment 1 circumvented this problem by adding a finite declarative clause, in which complementizers are optional (e.g., "you thought [(that)…"), inside an island (resulting in sentences like: "…who I wondered [why you thought [(that) he attacked __]]"). The experiment was an online acceptability judgment task with a high sample size and a standard experimental design (i.e., including familiar features like randomization and counterbalancing).

If subject gaps are inherently worse than object gaps in islands, then subject gaps should be rated worse than object gaps regardless of whether they follow a null (∅) or overt (*that*) complementizer. But if the That-Trace Effect and the Island Boundary-Gap Effect reflect the same constraint, then subject gaps should be worse than object gaps following the overt complementizer, but not when following the null complementizer.

2.1 Method

2.1.1 Participants

As the acceptability of the critical stimuli was expected to be quite low on the whole, a high target N of 150 was set to mitigate concerns about potential floor effects. To achieve this goal, 200 slots were made available online to UC San Diego undergraduates; a total of 161 participated for course credit.

All participants were native monolingual speakers of American English (defined as not having learned any other language before the age of 7) and were at least 18 years old. Participants were excluded if they answered fewer than 60% of comprehension questions correctly (50 participants), or if they did not rate "ceiling" fillers – designed to establish an acceptability ceiling (see below) – more than two Likert-scale points higher on average than "floor" fillers – designed to establish an acceptability floor (30 participants, including 25 who were also among the 50 reported with below 60% accuracy). A total of 106 participants were included in the analysis.[3]

2.1.2 Procedure

The procedure was the same for all experiments, which were conducted online using the Ibex Farm server (Drummond 2013). Participants began by providing informed consent. Subsequent instructions explained that they would be asked to rate the acceptability of sentences based not on the types of formal grammatical principles they may have learned in high school, but on subjective intuitions. They were told to consider whether, given an appropriate context, a native speaker of English might ever say such a sentence. Sentences appeared in full, with no time limit, in the center of the screen above a 9-point Likert scale with 1 labeled "bad" and 9 "good." After two practice trials, participants proceeded to the task. After every critical trial and most filler trials, participants responded to a multiple-choice comprehension question. Both Experiments 1 and 2 took roughly 20 minutes to complete.

2.1.3 Materials

Two factors were manipulated in a fully crossed 2×2 design. The first, COMPLEMENTIZER, had levels *null* and *that*. The second, GAP POSITION, had levels *subject* and *object*.

Critical trials consisted of 24 item sets, an example of which is given in Table 2. Stimuli contained relative clauses with two clausal embeddings. The first embedding ("…I wondered [why…") introduced an island. Inside the island, the second embedding introduced a finite declarative clause ("…you thought [{∅,that}…"). This clause contained a gap in either subject or object position. Because it was finite and declarative, the complementizer was optional. This allowed for the manipulation of the presence of a clause-initial function word inside an island.

An additional 24 fillers were randomly mixed with the critical items. These included 12 *ceiling* items, such as "Somebody mentioned that you liked wine so I brought a bottle of Chardonnay." *Ceiling* fillers were intended to be unambiguously highly acceptable

[3] Note that this reflects an unusually high number of participant exclusions. This was deemed acceptable in all experiments because the exclusion criteria seemed a priori reasonable, and even without these participants the experiments still had more participants than the typical linguistic experiment. Post-hoc analyses revealed that critical findings are robust to various exclusion criteria; original data are available online at https://osf.io/fn5at/.

Table 2: Experiment 1 stimuli. Sentences appeared in a 2×2 design. The GAP POSITION manipulation is shown in rows; COMPLEMENTIZER is manipulated in-line. Clause boundaries (square brackets) and gaps (underscores) were not visible to participants.

GAP POSITION	STIMULUS
Subject	It was the doctor [who I wondered [why you thought [{∅, that} __ hit the lawyer with a bat]]].
Object	It was the doctor [who I wondered [why you thought [{∅, that} the lawyer hit __ with a bat]]].
COMPREHENSION QUESTION	What happened? (a) I hit the lawyer with a bat. (b) The lawyer hit the doctor with a bat. (c) The doctor hit the lawyer with a bat. (d) You hit the doctor with a bat.

so as to establish a ceiling for acceptability ratings. The remaining 12 *floor* fillers were intended to establish a lower bound for acceptability ratings in hopes of pushing the ratings of the critical items up to avoid a floor effect. These consisted of doubly-center-embedded relative clauses, such as "This park that the landscaper that the architect hired revitalized attracted tourists."

Stimulus order was pseudo-randomized such that every consecutive group of four critical trials contained one instance of each critical condition. Four lists were automatically generated by the Ibex Farm software according to a Latin square design, and participants were assigned to lists cyclically in the order in which they began the experiment. The software was designed so that the particular list the next participant was assigned to was incremented only once the previous participant completed the consent form. This meant that participants who began the experiment before the previous participant completed the consent form were assigned to the same list as the previous participant, resulting in a slightly uneven distribution across lists. The lowest number of participants in a given list was 35 and the highest was 43.

2.1.4 Trial exclusions and data preparation

For Experiments 1 and 2, trials completed in less time than would be required to read the sentence at a rate of one word per 150 ms were excluded (25 critical trials, or 0.98%). Critical trials were also excluded if the comprehension question was answered incorrectly (566 trials, or 22.25%).[4] The remaining 1967 critical trials were included in the analysis; the number of observations per cell ranged from 440 to 533.

[4] As with any analysis, the approach outlined here includes a number of experimenter degrees of freedom, including the choice to exclude trials with incorrect responses. A series of supplementary analyses, reported in the appendices, show that the patterns of results across experiments were relatively robust to these choices. Appendix A reports the results of a model analyzing data from Experiments 1 and 2 in which trials were not excluded on the basis of an incorrect comprehension question response if the stimulus involved an island violation.

Raw ratings were z-scored by participant after trial exclusions (e.g., Schütze & Sprouse 2014). This served to reduce potential individual differences in the use of the Likert scale, as well as to increase the degree to which model residuals approximate a normal distribution.

2.1.5 Analysis

A linear mixed effects regression was used to model z-scores as a function of COMPLEMENTIZER, and GAP POSITION, and their interaction (Bates et al. 2015; R Core Team 2013). Both fixed effects were treatment coded. With treatment coding, the model terms reflect differences between particular conditions and a "baseline" condition – the model intercept. Thus, treatment coding is useful in cases where the researcher wants to compare conditions. Here, the intercept was the *That,Subject* condition and treatment coding was used to make three comparisons. The simple main effect of COMPLEMENTIZER reflects the difference between the *Null,Subject* condition and the baseline condition. The simple main effect of GAP POSITION reflects the difference between the *That,Object* condition and the baseline condition. Finally, the interaction term reflects the difference between the observed value for the *Null,Object* condition and what would be expected based on an additive combination of the two simple main effects.

Random intercepts for items and participants were included, and all fixed effects were allowed to vary within levels of the random effects. The model converged with the full random effects structure. The *summary()* function from the *lmerTest* package was used to obtain p-values (Kuznetsova et al. 2017). To confirm that critical effects significantly contributed to model fit, χ^2 tests were performed to compare models with and without the effect. All data and analyses are available online at https://osf.io/fn5at/.

2.2 Results

Results (summarized in Table 3 and depicted in Figure 1) showed a significant simple main effect of COMPLEMENTIZER. Because the factors were treatment coded, this effect reflects the fact that the *Subject,Null* condition was better than the *Subject,That* condition ($t(76.51) = 4.55$, $p < .001$); model comparison confirmed the effect significantly contributed to model fit ($\chi^2(1) = 18.185$, $p < .001$). There was also a significant simple main effect of GAP POSITION, reflecting the fact that the *object,that* condition was better than the *subject,that* condition ($t(399.78) = 2.2$, $p = .03$; model comparison: $\chi^2(1) = 4.839$, $p = .028$). Critically, the interaction was significant, reflecting the fact that in islands, subject gaps are only worse than object gaps in the presence of an overt complementizer ($t(135.628) = -3.15$, $p < .01$; model comparison: $\chi^2(1) = 9.803$, $p = .002$).[5]

Table 3: Experiment 1 results.

	β	D.F.	t	p	
Intercept (*That,Subject*)	−0.426	315.598	−17.285	< .001	***
COMPLEMENTIZER (*Null*)	0.161	76.508	4.552	< .001	***
GAP POSITION (*Object*)	0.071	399.785	2.202	.028	*
COMPLEMENTIZER × GAP POSITION	−0.134	135.628	−3.150	.002	**

[5] Unexpectedly, ratings of the *That,Object* condition appeared to be lower than those of the *Null,Object* condition. However, pairwise comparisons revealed no significant difference between the two conditions (*That,Object* − *Null,Object* = −0.027, $t(47.1) = -0.700$, $p = .897$), suggesting that this numerical difference was just noise.

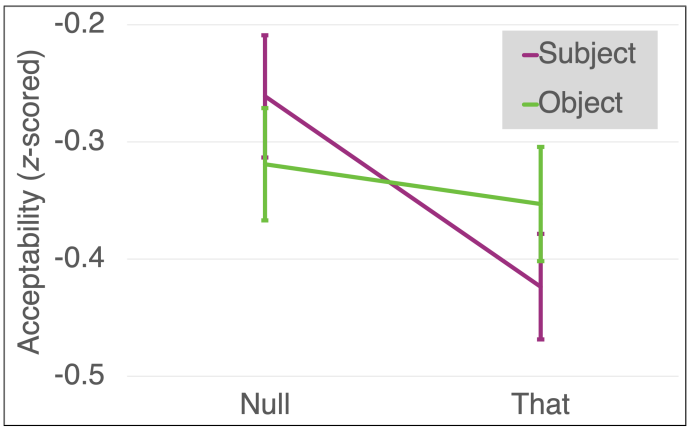


Figure 1: Experiment 1 results: Mean z-scores per condtion. Bars denote standard error.

2.3 Discussion

Experiment 1 tested the hypothesis that the Island Boundary-Gap Effect derives from a specific constraint on subject gaps in islands, in which case it would reflect a different constraint from whatever gives rise to the That-Trace Effect. This hypothesis would predict that subject gaps are always less acceptable than the corresponding object gaps in islands. However, the results contradict this prediction. Instead, subject gaps are only worse when immediately preceded by a clause-initial function word. This is consistent with the idea that the Island Boundary-Gap Effect may be the same as the That-Trace Effect.

One potential loose end is that the Experiment 1 stimuli were atypical of stimuli commonly tested in island literature in that they had an extra level of embedding. In order to confirm that the findings of Experiment 1 can be brought to bear on the question at hand, it should be verified that these stimuli do in fact behave like islands. This was the aim of Experiment 2.

3 Experiment 2

In the experimental literature, islands have been studied by systematically isolating component parts of the stimuli so as to parcel out the specific effect of an island violation on acceptability (Sprouse 2007; Sprouse et al. 2016). For instance, one can isolate the effect of an embedded gap on acceptability by comparing ratings of sentences like (5), which has a *matrix* (or non-embedded) gap, with sentences like (6), which has an embedded gap. Similarly, by comparing (5), which is a non-island structure, to (7), which involves an island structure, one can isolate the effect of islandhood.

(5) *Non-embedded gap:*
That's the judge who __ knew [that the defendant blackmailed the juror].

(6) *Embedded gap:*
That's the juror who the judge knew [that the defendant blackmailed __].

(7) *Non-embedded gap + island (no violation):*
That's the judge who __ knew [whether the defendant blackmailed the juror].

In this literature, the defining feature of an island is that its acceptability combines superadditively with the acceptability of an embedded gap. Thus, when an embedded gap appears inside an island structure, as in (8), the resulting acceptability is not simply the acceptability of a non-island structure (5) minus penalties for an embedded gap (i.e., (6) − (5)) and an island structure (i.e., (7) − (5)). Instead, it is even lower.

(8) *Embedded gap + island (violation):*
*That's the juror who the judge knew [whether the defendant blackmailed __].

This superadditivity is typically modeled as the statistical interaction between (a) whether a clause is an island or a non-island, and (b) whether a gap appears inside the clause or outside the clause (e.g., Sprouse et al. 2012; 2016). If the stimuli tested in Experiment 1 behave like typical islands, then they should show this characteristic interaction. Experiment 2 tested this prediction with another acceptability judgment study.

3.1 Method

3.1.1 Participants

An a priori target of 150 participants was set; 200 online experiment slots were made available to UC San Diego undergraduates and a total of 189 participated for course credit. Exclusion criteria were the same as in Experiment 1: participants had to answer more than 60% of comprehension questions correctly (50 exclusions) and/or rating ceiling fillers more than two points higher on average than floor fillers (14 partially-overlapping exclusions). A total of 131 participants were included in the analysis. None had participated in Experiment 1.

3.1.2 Materials

Two factors were manipulated, resulting in a fully crossed 2 × 2 design. The first factor, ISLANDHOOD, had levels *non-island* and *island*. The second, GAP EMBEDDING, had levels *matrix* and *embedded*.

Materials, demonstrated in Table 4, were adapted from those in Experiment 1.[6] They had the same global structure and same lexicalizations (e.g., "*the doctor said,*" "*you*

[6] Due to an oversight, the design of the stimuli was slightly different from that in Experiment 1. It is unlikely that this difference affected the results in any important way, but in the interest of transparency it is explained here. In Experiment 1, stimuli were designed so that the head noun was the same across all four conditions within a given item (i.e., all four conditions for the first item began with "It was the doctor who;" see Table 2). The trade-off for keeping the head noun consistent across conditions was that the meaning of the proposition in the lowest clause differed across conditions as a function of the GAP POSITION manipulation. In the *subject* gap condition ("(that) __ hit the lawyer with a bat"), the doctor is the agent of the hitting event. In the *object* gap condition ("(that) the lawyer hit __ with a bat"), the doctor is the patient. In contrast, in Experiment 2, stimuli were designed so that the event remained the same across all four conditions. However, this meant that the head noun differed as a function of the GAP POSITION manipulation (see Table 4). Fortunately, this seems unlikely to impact the generalizability of Experiment 2 findings to Experiment 1 given that the global structure of the stimuli was the same across experiments. One might be concerned that changing the role of the gap across conditions in Experiment 1 would result in different conditions having different degrees of plausibility. However, the stimuli were designed so that any ordering of the nouns in the sentence was roughly equally (im)plausible in order to ensure participants had to parse the dependency to correctly respond to the comprehension question (rather than reasoning over world knowledge). That is, there were no sentences with events like *lawyer defends burglar*, where *lawyer* is a priori much more likely to be the agent of the verb *defend* than *burglar*, so the plausibility of the events remained relatively constant across conditions.

thought," etc.), but differed in the particular factors manipulated. All conditions contained subject gaps, but the subject gap was either in a matrix or an embedded clause. The null complementizer was always used immediately before the embedded gap to avoid the That-Trace Effect. The same 24 fillers from Experiment 1 were used.

Table 4: Experiment 2 stimuli. Sentences appeared in a 2×2 design. The GAP EMBEDDING and ISLANDHOOD manipulations are shown in rows. Clause boundaries (square brackets) and gaps (underscores) were not visible to participants.

GAP	ISLANDHOOD	STIMULUS
Matrix	*Non-island*	It was the doctor [who __ said [that you thought [the lawyer hit me with a bat]]].
	Island	It was the doctor [who __ wondered [why you thought [the lawyer hit me with a bat]]].
Embedded	*Non-island*	It was the lawyer [who the doctor said [that you thought [__ hit me with a bat]]].
	Island	It was the lawyer [who the doctor wondered [why you thought [__ hit me with a bat]]].

Stimulus order randomization and list assignment followed the same procedure outlined for Experiment 1. The lowest number of participants in a given list was 44 and the highest was 51.

3.1.3 Trial exclusions and data preparation

A total of 820 critical trials (26.08%) were excluded: 814 for having the comprehension question answered incorrectly and a partially overlapping set of 15 for being read too quickly. The remaining 2324 critical trials were included in the analysis; the number of observations per cell ranged from 547 to 628. All other details were the same as in Experiment 1.

3.1.4 Analysis

All details of the analysis were identical to those in Experiment 1.

3.2 Results

Results, summarized in Table 5 and depicted in Figure 2, showed a significant simple main effect of ISLANDHOOD, reflecting the fact that sentences with islands were better than sentences without islands when the gap appeared in the matrix position (i.e., outside the island; $t(57.58) = 2.27$, $p = .03$); model comparison confirmed that the effect significantly contributed to model fit ($\chi^2(1) = 5.064$, $p = .024$). There was also a significant

simple main effect of GAP EMBEDDING, reflecting the fact that in non-islands, embedded gaps were less acceptable than matrix gaps ($t(66.25) = -11.57$, $p < .001$; model comparison: $\chi^2(1) = 55.868$, $p < .001$). Crucially, the characteristic island interaction was also significant, indicating that the difference in acceptability between a gap in an embedded versus a matrix clause is larger for island sentences than for non-island sentences ($t(604.4) = -3.73$, $p < .001$; model comparison: $\chi^2(1) = 13.068$, $p < .001$).

Table 5: Experiment 2 results.

	β	D.F.	t	p	
Intercept (*Non-island,Matrix*)	0.054	41.076	1.333	.190	
ISLANDHOOD (*Island*)	0.096	57.584	2.267	.027	*
GAP POSITION (*Embedded*)	−0.453	66.249	−11.569	< .001	***
ISLANDHOOD × GAP POSITION	−0.194	604.399	−3.733	< .001	***

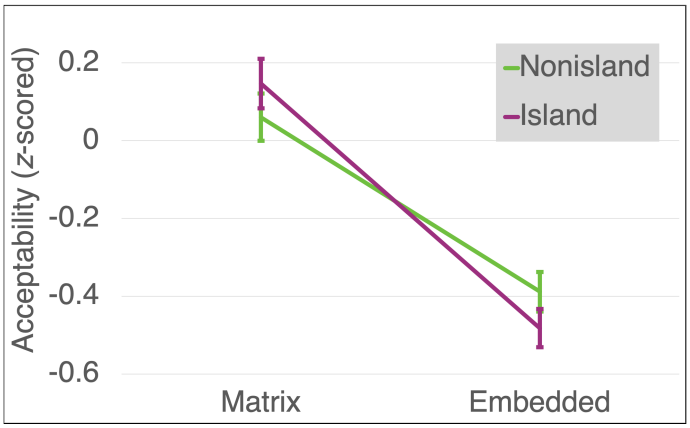


Figure 2: Experiment 2 results: Mean z-scores per condition. Bars denote standard error.

3.3 Discussion

Experiment 2 demonstrates that the sentences tested in Experiment 1 do in fact behave like islands, showing the defining statistical interaction between ISLANDHOOD and GAP EMBEDDING. However, two aspects of the current findings are atypical with respect to the commonly found pattern in the literature. First, when gaps were in matrix position, the *Island* condition was rated higher than the *Non-island* condition. Second, while significant, the difference between the *Non-island,Embedded* condition and the *Island,Embedded* condition was not as big as is commonly observed (see, for instance, the many results in (Sprouse et al. 2016)).

The first atypicality – that islands were more acceptable than non-islands in *Matrix* conditions – may reflect the increased syntactic diversity of the clause types in the sentence. It has previously been demonstrated that increasing the heterogeneity of syntactic properties of complex sentences can increase their acceptability. For instance, (9a) contains three full Determiner Phrases (*the mouse*, *the cat*, and *the dog*), and is highly unacceptable. However, (9b), which has the same global structure but contains one quantified nominal (*someone*), one local pronoun (*I*), and one full Determiner Phrase (*the mouse*), is much

more acceptable. This is thought to be the result of the fact that mutual distinctiveness among aspects of a stimulus can facilitate working memory processes (Bever 1974).

(9) a. *Similar nominals*
*The mouse [that the cat [the dog chased] stalked] squeaked.
b. *Distinctive nominals*
The mouse [that someone [I know] adopted] squeaked.

In the Experiment 2 stimuli, the types of nominal elements did not vary systematically across conditions as in (9), but the types of clause structures did. In *Island* conditions, stimuli consisted of two declarative clauses and one embedded interrogative clause, whereas *Non-island* stimuli consisted of three declarative clauses. It stands to reason that in these highly demanding sentences, structural diversity may have facilitated working memory in the same way that nominal distinctiveness aids working memory in (9), resulting in the higher acceptability of the *Island,Matrix* condition relative to the *Non-island,Matrix* condition.

An anonymous reviewer pointed out the second atypical feature of the Experiment 2 results: that the difference between the two embedded conditions was smaller than is often reported in the literature. This is potentially concerning because, if the size of this difference is indeed smaller than for the types of island structures that have been previously studied, it may mean that these stimuli are somehow different from "ordinary" island violations. If that were true, then the conclusion of Experiment 1 – that in islands, subject gaps are not inherently worse than object gaps – may not extend to islands in general.

However, there are at least three reasons not to be overly concerned about this discrepancy. First, if the structural diversity explanation given above is correct, then the *Island,Embedded* condition is worse than the *Non-island,Embedded* condition *in spite of* an acceptability boost the former receives from containing mutually distinctive clause types. Without this boost, the difference would be bigger.

Second, the absolute size of a difference in z-scores should not generally be compared across experiments. This is because a unit in z-scores represents one standard deviation in the distribution of ratings of all stimuli, and therefore depends on the spread of acceptabilities in the particular stimulus set. Thus, in a hypothetical experiment with the same critical items but different fillers, the size of this difference could be larger or smaller.

Third, and most importantly, the ratings in the *Island,Embedded* condition exhibit signs of a floor effect. That is, differences that we might have expected to be bigger are difficult or impossible to detect because there simply wasn't enough room at the bottom of the Likert scale to distinguish between, e.g., *very unacceptable* and *very, very unacceptable*. For instance, the *floor* fillers were included to establish a floor in the study. Thus, if the *floor* fillers were rated lower than the *Island,Embedded* fillers, then it suggests there *was* in fact room to distinguish even lower degrees of acceptability. But this was not the case. The mean z-score for *floor* stimuli was -0.521(S.E.: 0.055), and the mean z-score of the *Island,Embedded* condition was -0.482 (S.E.: 0.049). An uncorrected (i.e., liberal) t-test shows that the difference is not significant ($t(1056.1) = -1.331$, $p = .184$). Thus, if the *floor* items did in fact establish a floor in this experiment, then there is no evidence that the *Island,Embedded* condition was not also effectively at floor.

If the *Island,Embedded* condition was in fact confounded by a floor effect, this does not change the interpretation of the Experiment 2 results. In fact, any floor effect in this condition would only have made the interaction smaller, meaning that the critical

prediction was borne out *in spite of* a potential effect that worked against it, making the interaction even more credible.

Thus, despite these atypical features, the results of Experiment 2 verify that these stimuli behave like normal islands. This finding suggests that the conclusion of Experiment 1 – that subject gaps are not inherently worse than object gaps in islands – likely generalizes to the types of island structures more commonly tested in the experimental literature.

While the conclusion of Experiment 1 is consistent with the Island Boundary-Gap Effect and the That-Trace Effect resulting from the same underlying violation, it does not mean that this *must* be the case. Indeed, with a standard experimental approach such as those in Experiments 1 and 2, it is not generally possible to show that two constructs are the same in the Null Hypothesis Significance Testing framework.

As outlined in the introduction, one tactic for circumventing the inherent uninterpretability of a null result is to look for a relationship between individual differences in the That-Trace Effect and the Island Boundary-Gap Effect. In such an analysis, a significant result can constitute evidence for equivalence.

4 Experiment 3

One peculiarity of the That-Trace Effect is that it has long been reported that the effect varies considerably across speakers (Chomsky & Lasnik 1977; Pesetsky 1982; Sobin 1987; Featherston 2005). Experiment 3 takes a psychometric approach in order to measure this variability and compare it to variability in the Island Boundary-Gap Effect. If the That-Trace Effect and the Island Boundary-Gap Effect reflect the same constraint, and if there are meaningful individual differences in the strength of that constraint, then an individual who perceives the That-Trace Effect to be more unacceptable than the average English speaker should also perceive the Island Boundary-Gap Effect to be more unacceptable than the average English speaker. The effects should therefore correlate across individuals.

However, a difficulty with this type of design is identifying the source(s) of individual differences. Here, even if the two effects are not underlyingly the same, individual differences in their ratings are still likely to correlate. This is because a host of other processes that vary across individuals underlie behaviors like acceptability judgment. For instance, individual differences in pickiness may result in overall lower ratings on all sentences for some particularly highfalutin individuals. This could lead to a correlation that reflects individual differences in pickiness, but not in some single underlying linguistic constraint.

Thus, while one may surmise that if the two effects are underlyingly the same there should be a correlation in individual differences, the reverse inference – that a correlation is evidence of equivalence – is not necessarily true. In order to interpret a correlation as equivalence between two constructs, strong control measures must be in place so as to remove all other sources of covariance (Furr 2017; Crocker & Algina 1986). If such controls are adequately employed, as demonstrated below, then a correlation between individual differences in two variables may be regarded as evidence that the two share some processing machinery.

In the present study, one likely source of extraneous correlated individual differences is the shared syntactic makeup of the That-Trace Effect and Island Boundary-Gap Effect. For instance, both effects involve long-distance dependencies. If there are individual differences in the acceptability of long-distance dependencies, then these could drive the predicted relationship between individual differences in the That-Trace Effect and the

Island Boundary-Gap Effect. To mitigate this and similar concerns, two control conditions were included in the stimuli (see 4.1.2).

There are also nonlinguistic individual differences which could drive such a correlation. To minimize attentional differences, comprehension questions were included after every trial. To control for different uses of the Likert scale, control conditions were included in the analysis as covariates (see 4.2). And to avoid inadvertently inducing individual differences by exposing participants to different sentences in different orders, stimuli were not counterbalanced or randomized, meaning that each participant saw the exact same sentences in the exact same order, following common practice in psychometric studies (Swets et al. 2007).

4.1 Method

4.1.1 Participants

An a priori stopping point for running participants was set at either 150 participants or the end of the academic year at UC San Diego, whichever came first. Running was stopped at the end of the year with a total of 104 UC San Diego undergraduates having participated for course credit. None had participated in Experiments 1 or 2.

In Experiments 1 and 2, bilinguals had been excluded to avoid transfer effects, where aspects of one grammar bleed over into another (Gas 1979). If, for instance, the bilingual's other language does not have a That-Trace Effect, this might have reduced the size of the effect in Experiment 1. However, in Experiment 3 this was not a concern because, even if a bilingual's non-English grammar affects their English grammar, if the That-Trace Effect and the Island Boundary-Gap Effect are indeed the same, then whatever influence the other grammar exerts should be the same for both effects. (Indeed, to foreshadow, bilinguals' Experiment 3 data patterned with those of monolinguals.)

Speaking another language was therefore deemed orthogonal to the question of interest, so bilinguals were allowed to participate so long as they responded to questions in a demographics questionnaire indicating that they (a) self-identified as native English speakers, (b) learned English before the age of 7, and (c) were not aware of having any perceptible non-native accent. Of the 23 bilingual participants, 5 were excluded for not meeting these criteria. As in Experiments 1 and 2, participants were also excluded for answering fewer than 60% of comprehension questions correctly (4 exclusions) and/or for not rating ceiling fillers more than two points higher on average than floor fillers (5 exclusions). One additional participant was excluded due to data loss from trial exclusions (see below). A total of 93 participants were included in the analysis.

4.1.2 Materials

Materials, summarized in Table 6, consisted of two types of critical items (10 each), two types of control items (10 each), and three types of fillers (12 ceiling, 5 floor, and 10 "middle" fillers described below). Unlike in previous experiments, conditions were not counterbalanced across items. That is, items consisted of only one sentence rather than a set of four, so the particular lexicalizations in the stimuli were confounded with conditions.[7] Furthermore, the stimulus order was randomized just once, and then fixed

[7] Given no a priori reason to believe that properties like the particular tense/aspect or the use of modals might be able to drive the critical effect in this study, the stimuli were diversified with respect to these properties to make the task less monotonous. The sample stimuli in Table 6 are random samples from the full set, so, for instance, not all $Control_{TTE}$ conditions contained modals, and some items in other conditions did

so that all participants saw all stimuli in the same order. This removed the possibility that any observed individual differences might derive from differences in the particular sentences or the particular order that participants saw.

Table 6: Experiment 3: Sample stimuli.

CRITICAL CONDITIONS	
TTE	It was the one-legged pirate [who everyone was saying [that __ kidnapped the princess]].
IBGE	That's the royal [who everyone is wondering [whether __ will succeed the gravely ill king]].
CONTROL CONDITIONS	
Control$_{TTE}$	It must have been someone [who they think [∅ __ had access to sensitive information]].
Control$_{IBGE}$	It was the math professor [who the students all wondered [how you knew [∅ __ had a crush on the history professor]]].
FILLERS	
Ceiling	There was an enormous painting of a Teddy bear next to the fireplace.
Middle	The visitor's log from the secretary's secret desk drawer was taken out by Ming.
Floor	The bug that the kid that the teacher scolded trapped ran in circles.

The two critical conditions both contained subject gaps immediately following overt clause-initial function words. In the *TTE* condition, meant to measure the strength of the That-Trace Effect, the gap was in a non-island and the function word was *that*. In the *IBGE* condition, meant to measure the Island Boundary-Gap Effect, the gap was in an island and the function word was either *why* or *whether*.

The two control conditions matched the two critical conditions for as many features as possible, except for having a gap immediately after a clause-initial function word. These conditions were designed to remove various potential sources of individual differences, isolating the contribution of a gap immediately following a clause-initial function word. The *Control*$_{TTE}$ condition matched the *TTE* condition in everything except having an overt clause-initial function word. Similarly, the *Control*$_{IBGE}$ condition differed from the *IBGE* condition in that it did not have an overt clause-initial function word. However, in order to manipulate the presence/absence of this function word, an additional declarative clause had to be embedded in *Control*$_{IBGE}$ items, as in Experiment 1.

The extra level of embedding in *Control*$_{IBGE}$ conditions relative to *IBGE* conditions meant that these conditions differed in two ways. This additional difference is not ideal, but poses minimal risk to the interpretation of the results. The critical prediction is a correlation between (a) whatever individual differences exist in *TTE* stimuli above and beyond individual differences in *Control*$_{TTE}$ stimuli, and (b) whatever individual differences exist in *IBGE* stimuli above and beyond individual differences in *Control*$_{IBGE}$ stimuli. There is no reason to expect that individual differences in the acceptability of clausal embedding correlate with the strength of the That-Trace Effect. Therefore, any observed correlation

contain modals. In order for this to be able to drive a correlation between individual differences in *TTE* and *IBGE* above and beyond what can be accounted for by control conditions, differences in the use of modals (for example) would have to be correlated in the two critical conditions, but not in the control conditions. This was not the case.

can safely be assumed to derive from the only shared property of the two measures: the presence/absence of a gap immediately after a clause-initial function word.

Finally, an additional type of filler was added to this study because several participants in Experiments 1 and 2 reported anxiety about correctly responding to comprehension questions after *Floor* filler items, which were difficult to comprehend. The number of *Floor* items was therefore reduced to five, and ten "*Middle*" fillers were added. These were composed of an inanimate complex NP subject (e.g., "The cookies with dried rose petals and organic whole oats…") in a passive sentence with an explicit animate agent ("…were eaten by the customer."). Heavy NPs tend to be uttered later in English sentences (Ross 1967), and animate NPs tend to be uttered earlier (Bock & Warren 1985). These sentences therefore violated two norms, and were expected to be rated relatively low, but to be more acceptable and less stress-inducing than the *Floor* fillers.

4.2 Analysis

In addition to not randomizing trial order, two other psychometric practices were adopted. First, no trials were removed from the data prior to analysis. This is because excluding trials means that individuals' mean ratings reflect the means of different items, which could create artifactual individual differences.

Second, Experiment 3 analyzed raw ratings rather than z-scores. This is because z-scores remove individual differences by design. To calculate the z-score, the mean of all of a participant's ratings is subtracted from each of that participant's ratings to set each participant's mean rating to 0. Each of the participants' mean-adjusted ratings are then divided by the standard deviation of all of that participant's ratings, setting the standard deviation of each participant's ratings to 1. Typically, this is done to eliminate differences in the use of the Likert scale. While different uses of the scale are certainly a concern in Experiment 3, z-scoring can have the unintended consequence of reducing the individual differences we aim to study.

To see this, consider a participant who likes *TTE* stimuli more than the average person. For the sake of argument, let us also momentarily assume that there are no individual differences in the acceptability of the other conditions, nor in the use of the Likert scale. The participant who likes *TTE* stimuli more will have a higher overall mean rating (i.e., across all stimuli) than a participant who likes *TTE* stimuli less. In calculating these participants' z-scores, a higher mean will therefore be subtracted from the ratings of the participant who likes *TTE* stimuli more, and a lower mean will be subtracted from the ratings of the participant who likes the *TTE* less, thus reducing the difference between these two participants' ratings of *TTE* stimuli and obscuring the individual differences. Now, individual differences in other conditions and in the use of the Likert scale probably do exist, so the picture is somewhat noisier. But if these individual differences are not correlated with the individual differences in the *TTE* and *IBGE* (and there is no reason to expect that they are), then, across a large enough sample size, individual differences in other conditions and in the use of the Likert scale should cancel out, and the same logic holds.[8]

[8] There are still good arguments for z-scoring, even considering the concerns outlined here. One middle-ground approach may be to calculate z-scores based on ratings of filler items only, so that the individual differences of interest do not contribute to the participant means/standard deviations used to scale ratings. Analyses of ratings transformed in this way are reported in Appendix B. Results largely pattern with those of the main analyses, suggesting that the findings are relatively robust to experimenter degrees of freedom in data preprocessing.

To understand whether the That-Trace Effect and the Island Boundary-Gap Effect derive from the same constraint, it is not enough to show a correlation between participants' mean ratings of the two critical conditions, because a correlation could be driven by any number of extraneous things. Thus, rather than simply correlating ratings of these conditions with one another, each critical condition was modeled with a separate multiple linear regression in which the control conditions served as covariates. In Model 1, participants' mean ratings of *TTE* items were modeled as a function of participants' mean ratings of *IBGE* items, as well as $Control_{TTE}$ items and $Control_{IBGE}$ items. In Model 2, participants' mean ratings of *IBGE* items were modeled as a function of participants' mean ratings of *TTE* items, $Control_{IBGE}$ items, and $Control_{TTE}$ items.[9]

In the particular type of multiple regression employed here (R's default *lm()* function), multicollinearity (when two or more of the regressors are correlated) is dealt with by ignoring the shared variance (McElreath 2020; Ch. 6). Thus, by including the control conditions as covariates, it is ensured that any significant relationship between the two critical terms reflects individual differences in properties shared only by those two conditions.

This has a similar effect to simply subtracting each participant's mean rating of $Control_{TTE}$ stimuli from their mean rating of *TTE* stimuli and correlating that difference with the difference between *IBGE* and $Control_{IBGE}$ stimuli. While subtraction may be somewhat more intuitive, multiple regression offers a few additional benefits, including producing statistics indicating whether these "subtractions" remove meaningful variance.

To make this point more concrete, consider two individuals who map the same degree of subjective unacceptability to different points along the Likert scale. One may assign this degree of unacceptability a 4, and the other may assign it a 6. This baseline difference in ratings will likely mean that for each type of structure, the first person's ratings will be lower than the second person's. This alone is enough to drive a correlation between individual differences in two structures. One might therefore expect that *all* of the independent variables should significantly predict the dependent variables in Models 1 and 2.

However, because this variance is shared by all of the independent variables, the model cannot tell which of them to attribute it to, so it attributes it to none. Thus, individual differences that are common to all of a participant's acceptability ratings, such as those that might arise from differences in the use of the Likert scale, cannot drive a significant effect in the present analyses. If the *IBGE* term is significant in Model 1, it means that individual differences in the Island Boundary-Gap Effect correlate with individual differences in the That-Trace Effect above and beyond what can be explained by individual differences present in the ratings of $Control_{TTE}$ and/or $Control_{IBGE}$.

Another benefit of this analysis approach is that the models have built-in tests of two key concepts in psychometric research: *reliability*, or the accuracy of a measure, and *validity*, or the degree to which a measure reflects the construct it is meant to measure (Cronbach & Meehl 1955). Without high validity and reliability, individual differences are impossible to distinguish from naturally occurring noise.

The data can be said to have high reliability if a significant relationship is found between *TTE* and *IBGE* because this would indicate that acceptability ratings were an accu-

[9] Note that this approach does not take into account measurement error in the participant means. However, measurement error in independent variables leads to a systematic *under*-estimation of effect sizes and correspondingly higher p-values (Gorrod et al. 2013; Bound et al. 2001), thus making models that ignore this error (like the present ones) more conservative. One approach to accounting for measurement error here is demonstrated in Appendix C with Bayesian mixed-effects models.

rate enough measure to reveal individual differences in a shared underlying constraint, despite other differences between these two conditions.

If acceptability ratings have high enough validity, then the distinct properties of the two control conditions should lead to different patterns of significance in the two models. In Model 1, where the dependent variable is *TTE*, a significant effect of $Control_{TTE}$ is expected, but no effect of $Control_{IBGE}$ because, although it also shares some structural properties with *TTE*, all of these properties are also shared by $Control_{TTE}$. Similarly, in Model 2, $Control_{IBGE}$ is expected to be significant, but $Control_{TTE}$ is not because the former shares all of the same properties with *IBGE* as the latter.

One final difference between this analysis and that of Experiments 1 and 2 is that Experiment 3 employs a fixed-effects-only model. This is because there are no random effects to feed the model: random effects for participants are intended to remove individual differences, so these are not included, and random effects for items require that the model have access to item-specific data, but the present models analyzed aggregate statistics (i.e., participants' condition means).

4.3 Results

Mean ratings for each condition are shown in Figure 3. Because particular lexicalizations were not counterbalanced across items (i.e., the sentence about a pirate kidnapping a princess only ever appeared in the *Non-island,Overt* condition), any differences between conditions may reflect differences in the acceptability of the particular sentences in a condition. Comparing condition means therefore does not necessarily reveal anything about the true differences between the structures' acceptability, so no such statistical analysis was performed.

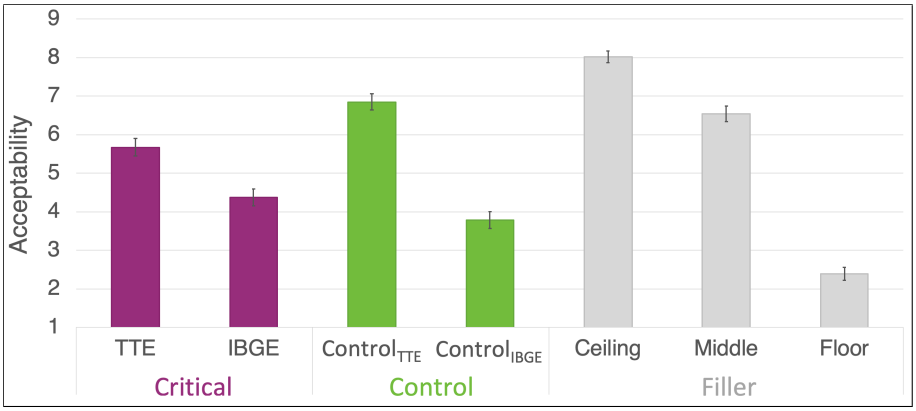


Figure 3: Experiment 3 results: Mean ratings per condition. Bars denote standard error.

Nonetheless, the pattern is what would be expected had the study included counterbalancing and randomization, perhaps indicating that any lexicalization-specific acceptability differences between conditions were relatively small. The *TTE* items were rated worse than the $Control_{TTE}$ stimuli, as expected due to the That-Trace Effect. In islands, this pattern reversed, and *IBGE* stimuli were rated better than $Control_{IBGE}$ stimuli. This is not surprising given that the $Control_{IBGE}$ condition contained one more clausal embedding than the *IBGE* condition, making it longer and more complex.

Results of both models are summarized in Table 7. Both revealed a strong positive relationship across participants between the strength of the That-Trace Effect and the Island Boundary-Gap Effect ($t(89) = 6.44$, $p < .001$; see Figure 4). Note that this is the

same effect in both models (and as such has the same t-value). The difference between Models 1 and 2 is in the relationships between the dependent variable and the control conditions. In Model 1, *TTE* was significantly predicted by $Control_{TTE}$ ($t(89) = 3.82$, $p < .001$), but not by $Control_{IBGE}$. In Model 2, the pattern was flipped, and *IBGE* was significantly predicted by $Control_{IBGE}$ ($t(89) = 7.562$, $p < .001$), but not by $Control_{TTE}$.[10]

Table 7: Experiment 3 results.

Model 1: TTE as function of IBGE, $Control_{TTE}$, & $Control_{IBGE}$

	Estimate	t	p	
Intercept	0.788	1.533	.139	
$IBGE$	0.631	6.439	$< .001$	***
$Control_{TTE}$	0.346	3.818	$< .001$	***
$Control_{IBGE}$	−0.065	−0.583	.562	

Model 2: IBGE as function of TTE, $Control_{TTE}$, & $Control_{IBGE}$

	Estimate	t	p	
Intercept	−0.299	−0.644	.521	
TTE	0.504	6.439	$< .001$	***
$Control_{TTE}$	−0.060	−0.686	.495	
$Control_{IBGE}$	0.588	7.562	$< .001$	***

4.4 Discussion

The main analysis revealed a significant relationship between *TTE* and *IBGE*. That is, a person who finds the That-Trace Effect to be especially bad is more likely to also find the Island Boundary-Gap Effect especially bad. This relationship was present even though the models removed variance associated with control conditions, indicating that it could not have been driven by some extraneous individual differences such as different uses of the Likert scale.

It also could not have been driven by individual differences in how people rate ungrammatical structures in general, because $Control_{IBGE}$, which is ungrammatical due to an island violation, did not significantly predict *TTE* in Model 1. The results therefore support the idea that the That-Trace Effect and the Island Boundary-Gap Effect reflect the same underlying violation.

5 General Discussion

This paper investigated a decades-old question: whether the Island Boundary-Gap Effect derives from the same underlying constraint as the That-Trace Effect. The two phenomena share surface similarities, but they differ in their underlying structures. Specifically, the offending word in the That-Trace Effect is a complementizer, *that*, while the offending word in the Island Boundary-Gap Effect can be a complementizer-like function word

[10] The *IBGE* and $Control_{IBGE}$ stimuli in Experiment 3 contained "why" and "whether" as relative pronouns introducing the island clause. An anonymous reviewer wondered whether these different types of wh-words might behave differently. Supplementary subset analyses, which can be found on the OSF repository, show that the pattern of significant results remains exactly the same for the models in Table 7 when run on subsets of the data with only the relativizer *why* or *whether*.

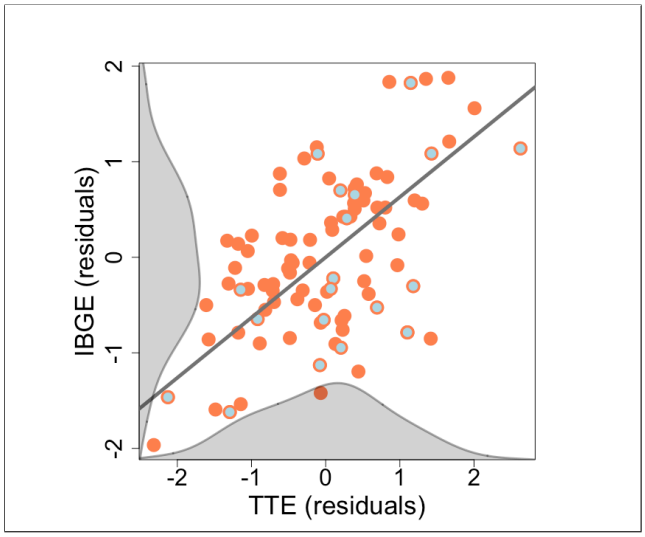


Figure 4: Experiment 3 results: Dots represent individuals; bilinguals have blue interiors. In order to show just the relationship between the That-Trace Effect and the Island Boundary-Gap Effect, the *TTE* and *IBGE* values here were both residualized on both control conditions. The dark grey line represents the best-fit line and density plots appear in light grey. A Pearson's correlation for the data in this plot reveals a strong, significant relationship: $r = .564$, $t(91) = 6.511$, $p < .001$).

(*because, after, wherever, …*), relative pronoun (*who, why, whether, etc.*), or even an entire referential phrase (*which of these staplers*). Three experiments found evidence supporting the idea that the Island Boundary-Gap Effect, at least as instantiated by *wh*-islands, shares a common cause with the That-Trace Effect.

Experiment 1 showed that the That-Trace Effect and the Island Boundary-Gap Effect are subject to the same licensing conditions. Specifically, it aimed to rule out the competing hypothesis that the Island Boundary-Gap Effect is the result of a specific constraint against subject gaps in islands. If instead the Island Boundary-Gap Effect has the same underlying cause as the That-Trace Effect, then subject gaps should not be less acceptable than object gaps in islands when they are not immediately preceded by a clause-initial function word. However, because clause-initial function words that introduce islands are never optional, subjecthood is almost always confounded with the position immediately after the clause-initial function word.

To remove this confound, the gap was embedded inside yet another clause, headed by the optional '*that*,' and the effect of a null vs. overt complementizer was assessed. Sentences with subject gaps were not rated worse than sentences with object gaps after a null complementizer, consistent with the idea that the Island Boundary-Gap Effect derives from whatever constraint also gives rise to the That-Trace Effect.

Experiment 2 validated the types of sentences used in Experiment 1 by demonstrating that they show the defining characteristic of islands: an interaction between islandhood and gap position.

The standard experimental approach thus demonstrated that the That-Trace Effect and the Island Boundary-Gap Effect are subject to the same distributional conditions. However, having similar licensing conditions is not proof that two constructs are the same. The standard approach leaves open the possibility that the two effects derive from different constraints, both of which happen to be alleviated in the presence of a null complementizer.

Experiment 3 employed psychometric methods to provide stronger evidence for equivalence. This experiment showed that individual differences in the acceptability of the Island Boundary-Gap Effect track with those of the That-Trace Effect, above and beyond any correlations with closely matched control structures. This indicates that the two effects share a common cause, at least with the Island-Boundary Gap Effect instantiated with *wh*-islands. It is left to future work to determine whether this finding holds for other types of islands.

5.1 Individual differences, reliability, and validity

In addition to demonstrating a novel approach to assessing equivalence between psychological constructs, Experiment 3 made another important contribution: the finding of individual differences in a syntactic representation. Contrary to the idealized notion that native speakers of the same language all have the exact same grammar, speakers in fact have slightly different grammars (Dąbrowska 2012). However, there is not much prior evidence for individual differences in syntax.

The clearest such data come from Han et al. (2007), who looked for evidence of different grammars among Korean speakers. Specifically, there are multiple syntactic positions that a verb might occupy due to the presence of a phenomenon known as *verb raising*. While in languages with verb-medial word order, like English, verb-raising results in a different surface word order, in verb-final languages like Korean, verb raising would not alter the surface word order. It is therefore possible that some Korean speakers might acquire a grammar with verb raising, while others acquire one without verb raising. Han et al. (2007) cleverly asked native Korean-speaking children and adults to interpret sentences with negation to determine whether the verb fell within the scope of negation, consistent with a non-verb raising grammar, or outside of the scope of negation, consistent with verb raising. Interestingly, results in both children and adults were bimodal: some speakers appeared to have acquired one grammar, while others had acquired a different one. (See also Dąbrowska 2008 for parallel work demonstrating bimodal individual differences in Polish morphology.)

One reason for the paucity of evidence for individual differences like these is the difficulty of controlling for confounds in this kind of research. For instance, individual differences in reading times in language comprehension tasks have been claimed to reflect individual differences in processing (Kidd et al. 2018). But it is trivially true that individuals' mean reading times will differ from one another given the presence of noise in the data. To credibly demonstrate individual differences, one must also demonstrate *reliability*. This was achieved in Experiment 3 by showing that individual differences correlated between the *TTE* and *IBGE* conditions.

But even when reliable individual differences are identified, it is often not possible to attribute them to any particular source. Reading time differences could reflect differences in language processing mechanisms, as Kidd et al. (2018) argue; but they could also reflect differences in visual processing, attention, or even different computer hardware (Enochson & Culbertson 2015). In addition to reliability, one must therefore also demonstrate that the measure is *valid*. Validity was demonstrated in Experiment 3 by showing that these conditions correlated with related control conditions, but not with unrelated ones.

The present work demonstrates a clear case of individual differences in native speakers' grammars. While individual differences in acceptability ratings are generally disregarded as noise, these findings show that they contain meaningful between-subjects variance.

Individuals in Experiment 3 differed not only in how strongly they dislike the That-Trace Effect/Island Boundary-Gap Effect, consistent with decades of speculation along these lines (Chomsky & Lasnik 1977; Pesetsky 1982; Sobin 1987), but also in how much they (dis)like island violations and ordinary relative clauses, as evidenced by the significant relationships between the critical items and control items.

5.2 Why is there a That-Trace Effect at all?

The finding that the That-Trace Effect and the Island Boundary-Gap Effect derive from the same underlying constraint makes an intriguing typological prediction: that languages should either show sensitivity to both effects or neither (see Goodall (2022) for similar reasoning about island phenomena). Indeed, this prediction appears to be borne out in Spanish (Torrego 1984; Suñer 1991) and Norwegian (Kush & Dahl 2020), both of which lack the That-Trace Effect and the Island Boundary-Gap Effect. A promising avenue for future work will be to establish whether this pattern holds for a wider set of languages. Such a finding would provide important validation for the unified analysis.

But a fundamental question still stands: Why should a constraint that prohibits gaps after clause-initial function words exist at all? And why in so many languages? One might posit that something about the semantic or pragmatic structure of these sentences is hard to process or ill-formed, as has been suggested for islands (Goldberg 1995). But this cannot be the case, as a sentence that is identical to one with a That-Trace Effect violation except in that it has no *that* is perfectly grammatical and has the same meaning.

This is all the more puzzling when one considers the fact that the existence of this constraint imposes a limit on the expressive capacity of a language. That is, in cases where clause-initial function words are obligatory (as in islands), English provides no good way to form a subject relative clause, as exemplified by (10):

(10) …the doctor who I wondered why __ was always running so late.[11]

As it is quite common to talk about doctors and why they are running late, it stands to reason that being able to express the intended meaning of this sentence, and similar ones, could be useful for speakers of languages that disallow it like English and Wolof. Why, then, should this particular structure be prohibited?

This study did not directly investigate the reason for the constraint, but the finding of individual differences in Experiment 3 may provide a clue. That is, whatever the underlying reason for the That-Trace Effect, it must be something that can vary across individuals. One possibility consistent with this is that the That-Trace Effect is an artifact of domain-general processing difficulty, as argued by McDaniel et al. (2015).

Another possibility is that the That-Trace Effect simply reflects the low frequency with which English speakers hear the violation (Phillips 2013). A child learning English might surmise that, if it were allowed in English, she would hear it more frequently. Because she doesn't, she implicitly learns that it is unacceptable, rates it as such when participating in a psycholinguistic experiment, and perpetuates this cycle by producing it infrequently. Children who happen to hear more of these structures early in life may go on to report that they are relatively more acceptable later in life, and children who hear fewer should show

[11] Of course, there is no *great* way of expressing this relative clause in the first place because of the island violation. But some islands, including *wh*-islands like the ones used throughout this paper, have been shown to be moderately acceptable with gaps in object positions, as in (8). The finding of Experiment 1 – that subject gaps are not worse than object gaps in islands when not immediately preceded by a clause-initial function word – suggests that, without the Island Boundary-Gap Effect, (10) might be a reasonably acceptable sentence.

greater sensitivity. This would be consistent with the fact that the constraint appears to vary across languages (Maling & Zaenen 1978; Nicolis & Biberauer 2008; Chacón 2015; Kush & Dahl 2020), and it converges with recent ideas that a similar process may be at least partially responsible for the idiosyncratic variation in which structures constitute islands across languages (Goodall 2019).

5.3 Causal inference

This paper has relied heavily on a correlation to infer a causal relationship. This inference is not generally licensed given that, if A correlates with B, it could be either because A causes B or because some latent variable, C, causes both A and B. Here it is assumed that a latent variable – maybe a constraint like (11) – causes both the That-Trace Effect and the Island Boundary-Gap Effect.

(11) A gap cannot immediately follow a clause-initial function word/phrase.

(See Bresnan 1972; 1977; Chomsky & Lasnik 1977 for similar proposals.) It is this constraint that is understood to vary across individuals, leading to the correlation between individual differences in the That-Trace Effect and the Island Boundary-Gap Effect.

In this case, the conclusion that a latent variable causes the correlation derives not from the data, but from reasoning about the problem. That is, it is not clear how the That-Trace Effect, a property of one type of sentence, could cause a distinct effect in different kinds of sentences. Thus, only the latent variable explanation makes sense.

In some cases, however, reasoning may not provide a clear answer. For instance, visual and verbal working memory appear to use partially overlapping resources (Morey & Cowan 2005). It is therefore possible that some component of visual working memory relies causally on the verbal working memory architecture (or vice versa). Here, however, the standard experimental approach can suffice to demonstrate both equivalence and causation. This is because working memory can be manipulated. For instance, if a researcher reduces verbal working memory capacity – for example by varying memory load, noise, attentional demands, or with transcranial magnetic brain stimulation – then it is relatively straightforward to infer the underlying relational architecture.

For linguistic constructs like the ones investigated here, it is not obvious how one might manipulate a syntactic "rule" like that in (11) to test for causal consequences. While the psychometric approach exemplified in Experiment 3 cannot definitively establish causality, a correlation may be enough in cases where causality can be inferred from the logic of the problem. For language research, then, the Experiment 3 approach may constitute a particularly useful tool.

6 Conclusion

In three experiments, this paper addressed a decades-old question in psycholinguistics: whether two ungrammatical constructions, the That-Trace Effect and the Island Boundary-Gap Effect, both derive from the same underlying constraint. It did so while contrasting two approaches to experimental design to demonstrate how researchers might use Null Hypothesis Significance Testing to demonstrate *equivalence* between psychological constructs.

Experiments 1 and 2 looked for evidence of equivalence using the standard framework, and demonstrate that the two effects are subject to the same distributional conditions: in both cases, acceptability is only reduced in the presence of an overt clause-initial

function word. While this finding is consistent with the idea that the two reflect the same constraint, it in no way requires this to be the case. Experiment 3 uses a psychometric approach and demonstrates that individual differences in the That-Trace Effect correlate with individual differences in the Island Boundary-Gap Effect, above and beyond what would be expected on the basis of individual differences in closely-matched structures. This would not be the case if the two did not share some underlying cause.

These experiments show that the That-Trace Effect and the Island Boundary-Gap Effect, at least as instantiated in *wh*-islands, derive from the same underlying constraint. The experiments also demonstrate that psychometrics is a powerful tool with which to address the often difficult question of equivalence within the Null Hypothesis Significance Testing framework. This approach may be particularly useful when dealing with constructs like grammatical constraints, which cannot easily be manipulated in a standard experimental paradigm.

Data Availability/Supplementary files

Data and analyses are publicly available online. https://doi.org/10.17605/OSF.IO/FN5AT

Ethics and consent

The experiments presented here were approved by the UC San Diego IRB. All participants provided informed consent prior to participation.

Funding information

This research was supported in part by the Graduate Research Fellowship Program at the National Science Foundation under grant DGE-1144086 to the author.

Acknowledgements

Thanks to Matt Wagers, Brianna Kaufman, and Victor S. Ferreira for invaluable help in the preparation of this manuscript.

Competing interests

The author has no competing interests to declare.

Authors' contributions

All authors contributed exactly equally.

References

Baltin, Mark. 2010. The nonreality of doubly filled comps. *Linguistic Inquiry* 41(2). 331–335. https://doi.org/10.1162/ling.2010.41.2.331.

Bates, Douglas, Martin Mächler, Ben Bolker & Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1). 1–48. https://doi.org/10.18637/jss.v067.i01.

Bever, Thomas G. 1974. The ascent of the specious; or, There's a lot we don't know about mirrors. In David Cohen (ed.), *Explaining linguistic phenomena*, 173–200. Washington, D.C.: John Wiley.

Bock, J. Kathryn & Richard K. Warren. 1985. Conceptual accessibility and syntactic structure in sentence formulation. *Cognition* 21(1). 47–67. https://doi.org/10.1016/0010-0277(85)90023-X.

Bound, John, Charles Brown & Nancy Mathiowetz. 2001. Measurement error in survey data. In James Heckman & Ed Learner (eds.), *Handbook of econometrics*, vol. 5, 3705–3843. New York: Springer-Verlag.

Bresnan, Joan W. 1972. *Theory of complementation in English syntax*. Cambridge, MA: MIT dissertation.

Bresnan, Joan W. 1977. Variables in the theory of transformations. In P. P. Culicover, T. Wasow & A. Akmajian (eds.), *Formal syntax*, New York: Academic Press.

Bürkner, Paul-Christian. 2017. brms: An R package for Bayesian multilevel models using Stan. *Journal of statistical software* 80(1). 1–28. https://doi.org/10.18637/jss.v080.i01.

Chacón, Dustin Alfonso. 2015. *Comparative psychosyntax*. College Park, MD: University of Maryland dissertation.

Chomsky, Noam. 1980. On binding. *Linguistic inquiry* 11(1). 1–46.

Chomsky, Noam. 1981. *Lectures on government and binding: The Pisa lectures*. Walter de Gruyter.

Chomsky, Noam. 2014. *The minimalist program*. MIT press.

Chomsky, Noam & Howard Lasnik. 1977. Filters and control. *Linguistic Inquiry* 8(3). 425–504.

Crocker, Linda & James Algina. 1986. *Introduction to classical and modern test theory*. ERIC.

Cronbach, Lee J. & Paul E. Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin* 52(4). 281. https://doi.org/10.1037/h0040957.

Dąbrowska, Ewa. 2008. The later development of an early-emerging system: The curious case of the polish genitive. https://doi.org/10.1515/LING.2008.021.

Dąbrowska, Ewa. 2012. Different speakers, different grammars: Individual differences in native language attainment. *Linguistic Approaches to Bilingualism* 2(3). 219–253. https://doi.org/10.1075/lab.2.3.01dab.

Drummond, Alex. 2013. Ibex farm. http://spellout.net/ibexfarm. Online server.

Enochson, Kelly & Jennifer Culbertson. 2015. Collecting psycholinguistic response time data using Amazon Mechanical Turk. *PloS One* 10(3). e0116946. https://doi.org/10.1371/journal.pone.0116946.

Featherston, Sam. 2005. That-trace in German. *Lingua* 115(9). 1277–1302. https://doi.org/10.1016/j.lingua.2004.04.001.

Ford, Marily. 1983. A method for obtaining measures of local parsing complexity throughout sentences. *Journal of Verbal Learning and Verbal Behavior* 22(2). 203–218.

Furr, R. Michael. 2017. *Psychometrics: An introduction*. Thousand Oaks, CA: Sage Publications.

Gas, Susan. 1979. Language transfer and universal grammatical relations. *Language learning* 29(2). 327–344.

Goldberg, Adele E. 1995. *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Goodall, Grant. 2019. Predicting the severity of island violations across languages: Some first steps. Talk presented at the 2019 LSA Annual Meeting session: Experimental approaches to cross-linguistic variation in island phenomena.

Goodall, Grant. 2022. D-linking, non-finiteness, and cross-linguistic variation in island phenomena. In *Theory and experiment in syntax*, Abingdon, U.K.: Routledge.

Gorrod, Emma J., Michael Bedward, David A. Keith & Murray V. Ellis. 2013. Systematic underestimation resulting from measurement error in score-based ecological indices. *Biological conservation* 157. 266–276. https://doi.org/10.1016/j.biocon.2012.09.002.

Granger, Clive W. J. 1969. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society* 424–438.

Hakes, David T., Judith S. Evans & Linda L. Brannon. 1976. Understanding sentences with relative clauses. *Memory & Cognition* 4(3). 283–290. https://doi.org/10.3758/BF03213177.

Han, Chung-hye, Noureddine Elouazizi, Christina Galeano, Emrah Görgülü, Nancy Hedberg, Jennifer Hinnell, Meghan Jeffrey, Kyeong-min Kim & Susannah Kirby. 2012. Processing strategies and resumptive pronouns in English. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*. 153–161. Cascadilla Proceedings Project Somerville, MA.

Han, Chung-hye, Jeffrey Lidz & Julien Musolino. 2007. V-raising and grammar competition in korean: Evidence from negation and quantifier scope. *Linguistic Inquiry* 38(1). 1–47. https://doi.org/10.1162/ling.2007.38.1.1.

Hawkins, John A. 2014. *Cross-linguistic variation and efficiency*. OUP Oxford.

Hoenig, John M. & Dennis M. Heisey. 2001. The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician* 55(1). 19–24. https://doi.org/10.1198/000313001300339897.

Holmes, Virginia M. & J. Kevin O'Regan. 1981. Eye fixation patterns during the reading of relative-clause sentences. *Journal of Verbal Learning and Verbal Behavior* 20(4). 417–430. https://doi.org/10.1016/S0022-5371(81)90533-8.

Kandybowicz, Jason. 2006. Comp-trace effects explained away. In Donald Baumer, David Montero & Michael Scanlon (eds.), *Proceedings of the 25th West Coast Conference on Formal Linguistics*, vol. 220. 228. Somerville, MA: Cascadilla Proceedings Project.

Kass, Robert E. & Adrian E. Raftery. 1995. Bayes factors. *Journal of the American Statistical Association* 90(430). 773–795.

Kayne, Richard S. 1981. ECP extensions. *Linguistic inquiry* 12(1). 93–133.

Keenan, Edward L & Bernard Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry* 8(1). 63–99.

Kenstowicz, Michael. 1989. The null subject parameter in modern Arabic dialects. In Osvaldo Jaeggli & Kenneth J. Safi (eds.), *The null subject parameter*, 263–275. Dordrecht: Kluwer.

Kidd, Evan, Seamus Donnelly & Morten H. Christiansen. 2018. Individual differences in language acquisition and processing. *Trends in Cognitive Sciences* 22(2). 154–169. https://doi.org/10.1016/j.tics.2017.11.006.

Kush, Dave & Anne Dahl. 2020. L2 transfer of l1 island-insensitivity: The case of norwegian. *Second Language Research* https://doi.org/10.1177/0267658320956704.

Kuznetsova, Alexandra, Per B. Brockhoff & Rune H. B. Christensen. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13). 1–26. https://doi.org/10.18637/jss.v082.i13.

Lasnik, Howard & Mamoru Saito. 1984. On the nature of proper government. *Linguistic inquiry* 15(2). 235–290.

Longo, Matthew R., Friederike Schüür, Marjolein P. M. Kammers, Manos Tsakiris & Patrick Haggard. 2008. What is embodiment? a psychometric approach. *Cognition* 107(3). 978–998. https://doi.org/10.1016/j.cognition.2007.12.004.

Maling, Joan & Annie Zaenen. 1978. The nonuniversality of a surface filter. *Linguistic Inquiry* 9(3). 475–497.

Martinović, Martina. 2013. Wh-morphology and cyclicity in Wolof.

McDaniel, Dana, Cecile McKee & Judy B. Bernstein. 1998. How children's relatives solve a problem for minimalism. *Language* 74(2). 308–334.

McDaniel, Dana, Cecile McKee, Wayne Cowart & Merrill F. Garrett. 2015. The role of the language production system in shaping grammars. *Language* 91(2). 415–441.

McElreath, Richard. 2020. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

Morey, Candice C. & Nelson Cowan. 2005. When do visual and verbal memories conflict? the importance of working-memory load and retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 31(4). 703–713. https://doi.org/10.1037/0278-7393.31.4.703.

Morgan, Adam Milton & Matthew W Wagers. 2018. English resumptive pronouns are more common where gaps are less acceptable. *Linguistic Inquiry* 49(4). 861–876. https://doi.org/10.1162/ling_a_00293.

Nicolis, Marco & Theresa Biberauer. 2008. The null subject parameter and correlating properties: The case of Creole languages. In Theresa Biberauer (ed.), *The limits of syntactic variation*, vol. 132, 271–294. Philadelphia, PA: John Benjamins.

Perlmutter, David M. 1968. *Deep and surface structure constraints in syntax*. Cambridge, MA: MIT dissertation.

Pesetsky, David. 1982. Complementizer-trace phenomena and the Nominative Island Condition. *The Linguistic Review* 1(3). 297–344. https://doi.org/10.1515/tlir.1982.1.3.297.

Phillips, Colin. 2013. On the nature of island constraints II: Language learning and innateness. In *Experimental syntax and island effects*, 132–157. Cambridge University Press.

R Core Team. 2013. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Riggs, Douglas S., Joseph A. Guarnieri & Sidney Addelman. 1978. Fitting straight lines when both variables are subject to error. *Life sciences* 22(1315). 1305–1360.

Rizzi, Luigi & Ur Shlonsky. 2008. Strategies of subject extraction. In *Interfaces + recursion = language?*, 115–160. Berlin: De Gruyter Mouton.

Rosner, B., D. Spiegelman & Walter C. Willett. 1992. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American journal of epidemiology* 136(11). 1400–1413.

Ross, John Robert. 1967. *Constraints on variables in syntax*. Cambridge, MA: MIT dissertation.

Rust, John & Susan Golombok. 2014. *Modern psychometrics: The science of psychological assessment*. Abingdon, U.K.: Routledge.

Schütze, Carson T & Jon Sprouse. 2014. Judgment data. In Robert J. Podešva & Devyani Sharma (eds.), *Research methods in linguistics*, vol. 27, 27–50. Cambridge University Press.

Sobin, Nicholas. 1987. The variable status of Comp-trace phenomena. *Natural Language & Linguistic Theory* 5(1). 33–60.
Sprouse, Jon. 2007. *A program for experimental syntax: Finding the relationship between acceptability and grammatical knowledge*. College Park, MD: University of Maryland dissertation.
Sprouse, Jon. 2009. Revisiting satiation: Evidence for an equalization response strategy. *Linguistic Inquiry* 40(2). 329–341. https://doi.org/10.1162/ling.2009.40.2.329.
Sprouse, Jon, Ivano Caponigro, Ciro Greco & Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in english and italian. *Natural Language & Linguistic Theory* 34(1). 307–344. https://doi.org/10.1007/s11049-015-9286-8.
Sprouse, Jon, Matt Wagers & Colin Phillips. 2012. A test of the relation between working-memory capacity and syntactic island effects. *Language* 88(1). 82–123. https://doi.org/10.1353/lan.2012.0004.
Suñer, Margarita. 1991. Indirect questions and the structure of CP: Some consequences. In Hector Campos & Fernando Martfnez-Gil (eds.), *Current studies in spanish linguistics*, 283–312. Washington, D.C.: Georgetown University Press.
Swets, Benjamin, Timothy Desmet, David Z Hambrick & Fernanda Ferreira. 2007. The role of working memory in syntactic ambiguity resolution: A psychometric approach. *Journal of Experimental Psychology: General* 136(1). 64–81. https://doi.org/10.1037/0096-3445.136.1.64.
Torrego, Esther. 1984. On inversion in spanish and some of its effects. *Linguistic inquiry* 15(1). 103–129.
van Gelderen, Elly. 2013. *Clause structure*. Cambridge University Press.
Wiener, Norbert. 1956. The theory of prediction. In *Modern mathematics for engineers*, New York: McGraw-Hill.

A Supplementary Analysis: Trial exclusion criteria in Experiments 1 and 2

An anonymous reviewer wondered whether excluding trials in which participants responded incorrectly to a comprehension question for a sentence with an island violation might be overzealous given that the stimuli were ungrammatical. The analyses in Experiments 1 and 2 were therefore repeated, but with different exclusion criteria.

For Experiment 1, participants were removed if they responded to fewer than 60% of the unambiguous *ceiling* fillers correctly (32 exclusions) or if they did not rate *ceiling* fillers more than two standard deviations higher than *floor* fillers, as before (8 more exclusions). Also as above, trials were removed if they were responded to in less time than it would take to read at a rate of 1 word per 150 ms (71 critical trials) and *ceiling* trials were removed if the comprehension question was responded to incorrectly (103 trials). No critical trials were excluded as all of these involved island violations. A total of 2833 remaining trials from 121 participants were analyzed (as compared to 1967 trials from 106 participants in the main analysis). As before, z-scores were computed by participant. All other analysis details were the same as in the main analysis.

Condition means are shown in Figure 5. Model results, reported in Table 8, were qualitatively identical to those reported in the main analysis in the paper (Table 3), suggesting that the Experiment 1 results were relatively robust to particular choices of exclusion criteria.

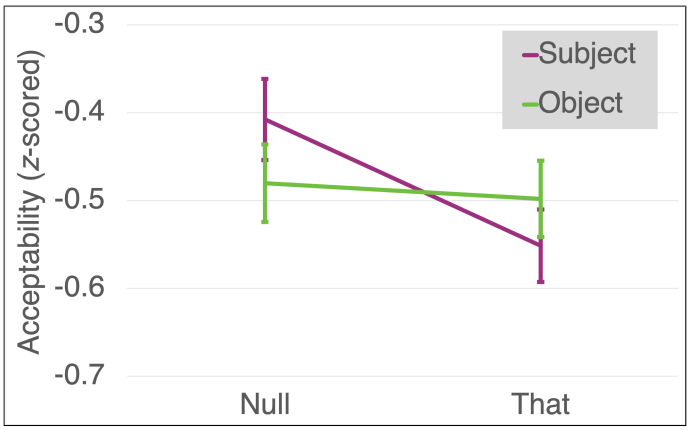


Figure 5: Experiment 1 condition means without excluding island violation trials for accuracy.

Table 8: Experiment 1: Results of supplementary analysis. The model was the same as the main model reported in the paper (see Table 3), but data included all trials that had previously been excluded because the comprehension question was answered incorrectly.

	β	D.F.	t	p	
Intercept (*That,Subject*)	−0.551	89.349	−24.427	< .001	***
COMPLEMENTIZER (*Null*)	0.143	55.393	4.938	< .001	***
GAP POSITION (*Object*)	0.054	549.764	2.085	.038	*
COMPLEMENTIZER × GAP POSITION	−0.126	99.399	−3.367	.001	**

For Experiment 2, participants were excluded if they responded correctly to fewer than 60% of comprehension questions to "unambiguous" stimuli (defined as *ceiling* fillers and the three critical conditions that did not involve an island violation: *Matrix,Non-island*, *Matrix,Island*, and *Embedded,Non-island*; 39 exclusions). Participants were also excluded if they did not rate *ceiling* fillers more than two standard deviations higher than *floor* fillers, as before (8 exclusions). Trials were excluded if they were read in less time than it would take to read at a rate of 1 word per 150 ms (20 critical trials). A total of 647 critical trials that did not involve island violations (19.1%) were excluded because the comprehension question was answered incorrectly. All other analysis details were the same as in the main analysis.

Condition means are shown in Figure 6. Again, model results, given in Table 9, were similar to those in the main analysis (Table 5), suggesting that the results of this experiment were also relatively robust.

B Supplementary Analysis: Analyzing z-scores in Experiment 3

The Experiment 3 models reported in Table 7 analyzed raw ratings. As explained in Section 4.2, z-scoring the ratings from Experiment 3 could have had the unintended consequence of reducing individual differences in *TTE* and *IBGE* stimuli. Specifically, participants who like *TTE* and *IBGE* stimuli more than the average person will likely have higher overall mean ratings on average. (*TTE* and *IBGE* stimuli made up roughly 30% of all stimuli in Experiment 3.) Subtracting these higher means from their ratings will reduce their ratings of *TTE* stimuli more than it would for a person who finds the

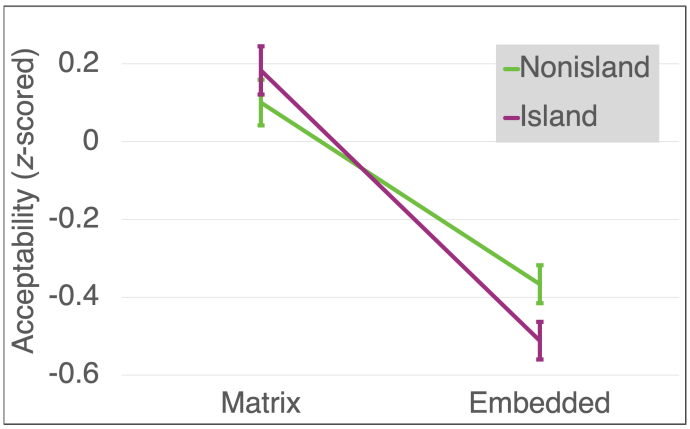


Figure 6: Experiment 2 condition means without excluding island violation trials for accuracy.

Table 9: Experiment 2: Results of supplementary analysis. The model was the same as the main model reported in the paper (see Table 5), but data included all island-violation trials that had previously been excluded because the comprehension question was answered incorrectly.

	β	D.F.	t	p	
Intercept (*Non-island,Matrix*)	0.093	35.781	2.313	.027	*
ISLANDHOOD (*Island*)	0.091	49.959	2.160	.036	*
GAP POSITION (*Embedded*)	−0.468	59.814	−12.070	< .001	***
ISLANDHOOD × GAP POSITION	−0.228	95.858	−4.577	< .001	***

TTE worse than the average person. Indeed, people who like *TTE* stimuli less will likely have lower overall mean ratings, and subtracting these means from their ratings will increase their ratings relative to participants who like *TTE* stimuli more. The end result is a reduction in the difference between these two types of participants, obscuring the individual differences of interest.

One potential way to avoid this would be to calculate z-scores with the means and standard deviations of just filler trials. In this way, participants' ratings of the critical and control conditions do not contribute to scaling their ratings. Such an approach comes with its own risks. For instance, participants' mean and standard deviations are calculated from only 27 ratings (just filler trials), meaning that these estimates are likely to be noisier than in an ordinary z-score approach (which would estimate these statistics using all 67 trials' ratings). The approach furthermore assumes that ratings of filler conditions are independent of ratings of critical and control conditions, which is probably not true. Sprouse (2009) demonstrated that participants in a binary forced choice task aim to give roughly the same number of each of the two types of response. It is easy to imagine that something similar happens with Likert scale rating tasks. For example, participants may aim to make their mean rating across stimuli a 5. If a participant strongly dislikes *TTE* stimuli, they may give fillers higher ratings to compensate for the lower ratings they assign to *TTE* items.

To determine whether anything hinged on the choice to analyze raw ratings in Experiment 3, the models reported in Table 7 were re-run using z-scores calculated using participants' means and standard deviations calculated using just filler trials. Results, reported in Table 10, largely pattern with those reported in the main analysis.

Critically, as in the main analyses, *IBGE* significantly predicts *TTE* in Model 1 and *TTE* significantly predicts *IBGE* in Model 2. Also as in the main analyses, $Control_{IBGE}$ z-scores significantly predicted *IBGE* z-scores in Model 2. The only difference in the pattern of significant/non-significant results between these analyses and the main ones reported in Table 7 is that $Control_{TTE}$ did not significantly predict *TTE* in Model 1. Note that this effect is not necessary for concluding that individual differences in the That-Trace Effect correlate with Individual Differences in the Island Boundary-Gap Effect. Thus, the results of this analysis suggest that the findings of Experiment 3 are relatively robust to Experimenter degrees of freedom in data preprocessing.

Table 10: Experiment 3: Results of supplementary analysis modeling modified z-scores.

Model 1: $TTE \sim IBGE + Control_{TTE} + Control_{IBGE}$

	Estimate	t	p	
Intercept	0.080	0.740	.461	
$IBGE$	0.539	5.884	$< .001$	***
$Control_{TTE}$	0.078	0.773	.442	
$Control_{IBGE}$	−0.033	−0.280	.780	

Model 2: $IBGE \sim TTE + Control_{TTE} + Control_{IBGE}$

	Estimate	t	p	
Intercept	−0.045	−0.420	.675	
TTE	0.519	5.884	$< .001$	***
$Control_{TTE}$	−0.074	−0.747	.457	
$Control_{IBGE}$	0.569	5.709	$< .001$	***

C Supplementary Analysis: Accounting for measurement error in Experiment 3

The Experiment 3 models analyzed participants' means in each of the critical and control conditions, thus ignoring measurement error associated with the means. There exist models, commonly used in biological and economic research, which aim to take into account the uncertainty in the predictors, giving better estimates of the slopes (βs). However, it is important to note that measurement error in the independent variable(s) affects model results in a different way from measurement error in the dependent variable, the latter of which is a much more familiar problem in psycholinguistics. With increased measurement error in the dependent variable, the result is increased uncertainty in the estimated slope, but not bias in any particular direction. However, when there is measurement error in the independent variable(s), as in the present case, it causes dilution (or attenuation bias): the systematic underestimation of the slopes and corresponding increase in p-values (Gorrod et al. 2013; Bound et al. 2001; Riggs et al. 1978; Rosner et al. 1992). Thus, if an accurate estimate of the effect size is the goal, then it is very important to take measurement error in the independent variables into account (although this is almost never done in psycholinguistics). However, if the goal is simply to establish a relationship, as in the present case, ignoring measurement error in the independent variables is

safe, and in fact a more conservative approach as the result will be an underestimation of the slope(s).

A clever suggestion by a reviewer was to analyze the Experiment 3 data by modeling individual differences as random effects in a linear mixed-effects model. The advantage of this approach is that it accounts for measurement error while simultaneously modeling item-based differences, thus removing another extraneous source of variance. A linear mixed effects model of raw ratings was fit with a four-level fixed effect term for CONDITION (levels: *TTE*, *IBGE*, $Control_{TTE}$, and $Control_{IBGE}$) with random effects for participants and items, and a random slope for condition nested within participants. The frequentist model would not converge, so the same model was fit in the Bayesian framework using the *brms* package (Bürkner 2017). The random effects matrix for participants was used as estimates of individual differences per condition.

These estimates were then plugged into the same simple linear models used in the manuscript (Table 7). The results, reported in Table 11, revealed a slightly different pattern of results from that reported in the paper. Critically, however, the relationship between the *TTE* and *IBGE* conditions remained positive and significant in both models.

Table 11: Experiment 3 results: Models of individual differences, as estimated by a Bayesian linear mixed-effects regression.

Model 1: $TTE \sim IBGE + Control_{TTE} + Control_{IBGE}$

	ESTIMATE	t	p	
Intercept	0.001	0.013	.990	
$IBGE$	3.624	4.685	$< .001$	***
$Control_{TTE}$	0.3688	1.118	$< .267$	
$Control_{IBGE}$	−2.678	−4.990	$< .001$	***

Model 2: $IBGE \sim TTE + Control_{TTE} + Control_{IBGE}$

	ESTIMATE	t	p	
Intercept	0.000	−0.062	.950	
TTE	0.055	4.685	$< .001$	***
$Control_{TTE}$	−0.330	−15.732	$< .001$	***
$Control_{IBGE}$	0.681	36.869	$< .001$	***