Using two-alternative forced choice tasks and Thurstone's Law of Comparative Judgments for Code-Switching Research

(In press, *Linguistic Approaches to Bilingualism*)

Hans Stadthagen-González*
(Corresponding Author)
Department of Psychology
University of Southern Mississippi
Hardy Hall 314
730 East Beach Blvd.
Long Beach, MS 39560
USA
h.stadthagen@usm.edu

Luis López
Department of Hispanic and Italian Studies
University of Illinois at Chicago
601 South Morgan Street (MC 315)
1722 University Hall
Chicago, IL 60607
USA
luislope@uic.edu

M. Carmen Parafita Couto
Leiden University Center for Linguistics
Witte Singel complex
Van Wijkplaats 3, 005B
2311 BX Leiden
Netherlands
m.parafita.couto@hum.leidenuniv.nl

C. Alejandro Párraga
Computer Vision Center,
Computer Science Department,
Universitat Autónoma de Barcelona,
Edifici 'O', Carrer de Les Sitges
Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallès) · Barcelona · Spain
Alejandro.Parraga@cvc.uab.es

Abstract

This article argues that 2-alternative forced choice tasks and Thurstone's law of comparative judgments (Thurstone, 1927) are well suited to investigate code-switching competence by means of acceptability judgments. We compare this method with commonly used Likert scale judgments and find that the 2-alternative forced choice task provides granular details that remain invisible in a Likert scale experiment. In order to compare and contrast both methods, we examined the syntactic phenomenon usually referred to as the *Adjacency Condition* (AC) (apud Stowell, 1981), which imposes a condition of adjacency between verb and object. Our interest in the AC comes from the fact that it is a subtle feature of English grammar which is absent in Spanish, and this provides an excellent springboard to create minimal code-switched pairs that allow us to formulate a clear research question that can be tested using both methods.

Key Words: two-alternative forced choice and Thurstone's law; code-switching; acceptability judgments

1. Introduction

The purpose of this article is to show that the application of Thurstone´s Law of Comparative Judgment (Thurstone, 1927) to the analysis of 2-alternative Forced Choice (2AFC) Tasks provides a robust method to extract and analyse acceptability judgments in general, and code-switching (henceforth CS) in particular. This method is considered the gold standard for collecting and interpreting subjective introspective data in most areas of behavioral research (e.g.: Cattelan, 2012; Montag, 2006; Parraga, 2015), but has been hitherto conspicuously absent from use in linguistic acceptability judgments. Although conceptually simple, the formal formulation of Thurstone's Law (David, 1988; Green & Swets, 1966; Thurstone, 1927; Torgenson, 1958) can be a bit daunting for those unfamiliar with it, so we also present a step-by-step, layman's explanation of how to conduct the analysis with the aim of facilitating its implementation by other researchers in the field.

In the remainder of the introduction we discuss the use of acceptability judgments in linguistic research, we then make the case for the use of two-alternative forced choice tasks, and subsequently explain the principles of Thurstone's law of comparative judgment and its application to the analysis of 2AFC data. Section 2 motivates the experiments: it presents the AC more formally, justifies using code-switching to investigate it and lays out the research question and derived hypotheses. Section 3 presents a Likert-scale acceptability judgment task and shows the limitations of this approach (Experiment 1). Section 4 uses Thurstone's method to approach the same problem (Experiment 2). A discussion section compares the outcomes of these two methods of obtaining acceptability judgments for the study of CS, as well as the implications of our results for the theory of grammatical dependencies.

1.1 Acceptability Judgments in Linguistic Research

Subjective judgments of acceptability or grammaticality have been a rich source of data for the formulation and evaluation of linguistic theories (e.g.: Chomsky, 1957, 1986; Cowart, 1996; Schütze, 2016), providing information about internal grammatical representations that would be difficult or impossible to obtain from other types of data (Schütze & Sprouse, 2013); acceptability judgments allow us to study constructions that occur seldom in spontaneous data, and to compare them under controlled conditions not available in a corpus. Specifically in the area of code-switching research, acceptability judgments have been used to study a wide variety of issues such as determiner choice and gender assignment in determiner-noun switches or the relative order of adjective and noun in code-switching (e.g.: Badiola, Delgado, Sande, & Stefanich, 2018; Parafita Couto, Deuchar & Fusser, 2015; Parafita Couto, Munarriz, Epelde, Deuchar & Oyharcabal, 2016, among many others). While the core objective of all acceptability tasks is to determine an informant's perception of the well-formedness of a particular construction (Schütze & Sprouse, 2013), there are several ways that have been used to achieve this goal (for a detailed discussion, see Schütze, 2016).

The simplest type of judgment consists of asking informants whether a particular sentence is acceptable or not, that is, a Yes/No judgment. This method is intuitive for participants, but can be unsatisfactory in many contexts because it lacks granularity to detect fine differences between conditions (Sorace & Keller, 2004). Furthermore, this type of task is prone to response bias so that extra-linguistic factors may shift the criterion or threshold (Cowart, 1996) between yes and no answers (e.g.: Bialystok, 1979). Additionally, it has relatively low statistical power (Sprouse & Almeida, 2011; Schütze & Sprouse, 2013).

Likert scales provide more gradual data that can be used to determine the size of the difference between conditions (Schütze & Sprouse, 2013). Furthermore, providing ratings on an *n*-point scale is an

easy and familiar task for most people nowadays, and indeed the use of Likert scales to elicit acceptability judgments is quite popular in the literature. Despite its widespread use, Likert scales are not without problems, mainly because the steps in the scale are pre-defined, which limits the granularity and the range of the scale. First, there is no certainty that participants treat the points on the scale as equidistant with regards to their internal representation of the acceptability continuum (Bard, Robertson, & Sorace, 1996; Fukuda, Goodall, Michel, & Beecher, 2012). For example, participants may treat the distance between two points at the extremes of a scale as larger (or smaller) than the distance between two points near the middle of the scale. Second, the scale may force participants to compress their judgments to fit the points on the scale given, even if they are not an adequate representation of the levels of acceptability they may wish to represent (Sprouse, 2011). Finally, Likert scales tax memory resources with large sets of items because participants must keep a running tally of their previous ratings in order to locate each new item within the same scale.

Bard et al. (1996) and Cowart (1996) propose using Magnitude Estimation (ME) for the collection of acceptability data. In ME, participants are presented with a *standard* sentence that is paired with an arbitrary value on the scale called a *modulus* (Bard et al., 1996; Stevens, 1956). They are meant to allocate ratings for the test sentences making reference to that pairing by using the modulus as a unit; that is, they should calculate the ratio between the acceptability of the test sentence and the standard sentence (a sentence that is twice as acceptable as the standard would generate a score that is twice the modulus). This operation is meant to generate a continuous ratio scale with an unrestricted number of values that allows for proportional comparisons between points on the scale. A scale of this type potentially offers higher accuracy ratings than those from Likert scales (Bard, et al., 1996; Cowart 1997; Featherston 2005a, 2005b). However, several studies (Bader & Häussler, 2010; Fukuda et al., 2012; Gigerenzer & Richter, 1990; Gigerenzer, Krauss, & Vitouch, 2004; Sprouse, 2011; Weskott & Fanselow, 2008, 2011) have

found that ME exhibited the same sensitivity as Yes/No and Likert tasks for acceptability judgments. It has also been called into question whether participants actually carry out the ratio task as indicated in the instructions (Featherston, 2008; Luce, 2002; Narens, 1996; Sprouse, Schütze & Almeida, 2013; Weskott & Fanselow, 2011). Sprouse (2007, 2011) proposes that participants performing ME tasks may not use the reference sentences as their unit of measurement in a proportional judgment but seem to perform a linear Likert-style rating ("more-" or "less acceptable"), just with the advantage of an open ended scale. Even proponents of ME recognize that it may be unsuitable for informants with weaker numeracy skills (Sorace, 2010). In summary, ME, from the point of view of the participant, is quite complex and it doesn't seem to offer many advantages over Likert-style ratings (Fukuda, et al., 2012).

In this paper, we advocate yet another method to extract acceptability judgments from language consultants. Our goal is to combine ease of use for participants with high statistical power and consequent granularity. Thus, we propose collecting acceptability data with a two-alternative forced choice task and the analysis of such data applying Thurstone's Law of Comparative Judgment.

1.2 Two-Alternative Forced Choice Judgements

In the simplest version of the 2AFC task, participants are presented with pairs of stimuli and must choose which item is more acceptable, with pairwise comparisons covering all possible contrasts between conditions. The 2AFC task, as we understand it here, must not be confused with forced categorical classifications between "acceptable" and "unacceptable" that are sometimes called "forced choices" but that correspond to the Yes/No task described before (e.g.: Serratrice, Sorace, Filiaci, & Baldo, 2009). The 2AFC offers several advantages: from the point of view of psychophysics, comparative judgments are considered easier (Nunnally, 1967) and more reliable (Mohan, 1977) than ratings. Additionally, paired comparisons do not require a memory component (necessary for rating new items along the same scale as

previous ones), and avoid possible shifts in the internal rating scale used by participants when new items are presented (Parraga, 2015). Sprouse (2011) shows that 2AFC judgments have higher statistical power than Yes/No, Likert, and ME tasks, making it particularly suitable for detecting differences between conditions (Gigerenzer & Richter, 1990; Gigerenzer et al., 2004; Sprouse & Almeida, 2011). The 2AFC task has been used to study linguistic phenomena both in monolinguals (e.g.: Sprouse & Almeida, 2011; Tikofsky & Reiff, 1970) and bilinguals (e.g.: Onar Valk, 2014; Sorace, 1996; Sorace, Serratrice, Filiaci, & Baldo, 2009), mostly by applying it to a comparison between just two conditions. Sometimes the term 2AFC has been used to refer to tasks that, strictly speaking, are not a two-alternative forced choice. For example, Sprouse and Almeida (2102, p. 4) make reference to Bard et al. (1996, p. 34) as support for the prevalence of 2AFC tasks in what they call "traditional methods". In turn, Bard et al. (1996) cite data from Haegeman (1991), but the method described there does not correspond to a pairwise comparison between all conditions, as we understand 2AFC. As for CS research, examples are vanishingly rare. Toribio (2001) is sometimes cited as an example of a forced-choice task, but this is not exact: in Toribio's experiment, participants were asked to *rate* two sentences in tandem, whereas what we propose is to ask them to *choose* one of them, a more straightforward kind of task.

We propose that the application of 2AFC tasks is particularly useful to analyse CS data because CS is often judged to be too fluid a form of linguistic knowledge to be amenable to study using grammaticality judgments; along these lines, consider the following quotation by Pieter Muysken: "clearly it is difficult if not impossible to rely on judgment data" (Muysken, 2000, p. 13). There seems to be two factors in reaching this conclusion. The first is that, in many communities, CS is stigmatized, and linguists have concluded that this negative attitude towards CS can affect acceptability judgment tasks and lead them to reject sentences that their linguistic systems would in fact generate (cf. Anderson, 2006; Giancaspro, 2013; Munarriz & Parafita Couto, 2014; Parafita Couto, Deuchar & Fusser, 2015). The

2AFC task allows us to circumvent this problem because participants are asked to compare one code-switched sentence against another, they are not asked to compare a code-switched sentence against an ideal grammatical value.

The second problem that linguists sometimes suggest is that CS itself is too malleable; within this view, CS would be a linguistic structure built on the fly, a performance phenomenon not subject to the regular restrictions that define a human language grammar - and therefore there would be no such thing as "acceptable" or "unacceptable" CS. This second concern should be put to rest by the time the reader reaches the end of this article.

In the psychophysics literature, the method of choice for analysing 2AFC data is derived from Thurstone's Law of Comparative Judgment (Bock & Jones, 1968; Cattelan, 2012; Engen, 1971; Jones & Thissen, 2007; Parraga, 2015; Reber, 1995), which places the results of multiple pairwise comparisons along a single interval scale that represents a one-dimensional quality. This scale provides a high degree of granularity, not available when forced choice data is subjected to other types of analyses. More importantly, the unit along that scale is the standard deviation of the distribution of responses for that particular set of data, so the scale is not pre-determined and provides an unrestricted scale with potentially infinite granularity, like ME, but with much higher statistical power (Sprouse & Almeida, 2011).

To summarize what we have so far: We argue that 2AFC in combination with Thurstone's law of comparative judgment has several advantages over competing methods of extracting and analysing acceptability CS judgments: it has greater statistical power (and therefore granularity) than its alternatives, it is a simple task for participants because it does not tax their memories or request that they hold abstract scales in their minds and provides a simple path to avoid judgments marred by prescriptivism. In the following paragraphs we provide some additional background on Thurstone's law and we introduce the experiments we will use to illustrate the usefulness of this method.

1.3 Thurstone's Law of Comparative Judgment

The fundamental concept behind Thurstone's law of comparative judgment (Thurstone, 1927; for further details see Bock and Jones, 1968; Edwards, 1957; Torgerson, 1958) is that the proportion of times a stimulus is judged as having more of a given attribute (e.g.: more beautiful, better formed, more acceptable, or any subjective attribute being measured) than another is related to the number of units separating the two sensations in a psychological scale that represents that quality. For example, if in half of the comparisons object B is judged more pleasant (or "acceptable") than object A, both objects are equally pleasant. If object C is judged more pleasant than object B in most comparisons between them, object C is likely to be the most pleasant object of the three. The probability of two different stimuli having exactly the same value on the judgment scale is considered to be extremely small, and thus no "tie" is allowed when making the pairwise judgment (David, 1988). The proportion of preferences for each condition in each pairwise comparison is normalized and converted to standard deviations (which is the unit of measure for Thurstone's scale), and the average for each condition is calculated. Finally, the scores are ranked from lowest to highest and the scale is linearly shifted so that, by definition, the lowest score becomes the origin of the scale.

The results of Thurstone's analysis must be interpreted within the context of signal detection theory (Cowart, 1996). Every real-world measurement is the result of the sum of a signal (the "ideal" magnitude being measured), and a random noise component that usually follows a normal distribution. Repeated instances of the same measurement yield a normal distribution centered on the likely "real" value of the magnitude, with measured values decreasing in probability as they get further from that central value (that is, large errors of judgment in either direction are less likely to occur than small errors). Thurstone's "score" provides the center or mean of such a normal distribution, which Thurstone called

"discriminal dispersions", for each condition. These central values can then be interpreted as scale values measured on an interval scale that represents a psychological continuum (in our case, an acceptability scale). An interval scale is one in which the origin (that is, its 0-value) is arbitrary but the distance between values on the scale is meaningful, so Thurstone's analysis does not only yield a hierarchy of choices, but also a meaningful measure of the *degree of difference* between values for each condition (Stevens, 1946). The unit of measurement along that scale is defined as one standard deviation of the distribution for that particular set of data. The distance between those means represents the participants' ability to discriminate one pattern from another, while the degree of overlap between normal curves indicates the likelihood that an "inconsistent" decision will be made (Brown & Peterson, 2009). In our case, an inconsistent decision is one in which an informant would pick the "wrong" sentence as more acceptable because of noise, which in the context of acceptability judgments can be the result of fatigue, practice, priming, social pressure, lack of attention, or any number of unknown factors. It is important to understand that the Thurstone measurement model is concerned with the ranking of non-physical entities such as "beauty" or "acceptability" which in turn define the "psychological continuum" (scale) where the comparisons are made. This psychological scale is an artificial construct which involves no assumption of a normal distribution in the physical world: it is defined (spaced off) so that the frequencies of the discrimination processes for any given stimulus form a normal distribution on it. The very nature of the stimuli means that these units do not have the same meaning as physical units such as "kilograms" of "gallons", which are linked to invariant physical magnitudes of objects. In other words, saying that the beauty of a sculpture is 0.20 and that of another is 0.5 does not have the same meaning as saying that the weight of a sculpture is 100 kg and the weight of another is 20 kg. One can, however, draw comparisons involving the relative distances between conditions in the psychological continuum (in our case indicating the acceptability of the sentences). For further details, Tsukida and Gupta (2011) offer a concise tutorial

on the analysis and interpretation of paired comparison data.

The general case of Thurstone's law is concerned with paired comparison data obtained from a single judge when only two judgments are allowed for each observation and is mathematically insoluble. In order to make it soluble, Thurstone introduced a series of assumptions, namely: (1) the formulation used for repeated judgments by a single observer is valid for a group of observers; (2) the variability in judgements between two comparison stimuli (also called "discriminal deviations") is uncorrelated and not "grossly different". There are five "cases" or variations of Thurstone's general formula, each responding to different assumptions for the judgments being analysed. In the most favorable case (namely case V), Thurstone assumed that the standard deviations of these judgements are equal, which greatly simplifies the law's formulation. In other words, these assumptions mean that the stimulus series is very homogeneous with no distracting attributes and that the "quality" perceived in one of the attributes has no influence on the "quality" perceived in its comparison specimen (Thurstone, 1927). Psychophysical research has shown that in most instances it is safe to make these assumptions since deviations are generally small and do not influence the results greatly.

In this study we apply Thurstone's law of comparative judgment case V to generate an interval scale based on comparisons of pairs of code-switched sentences that indicates not only a ranking of acceptability but also the relative distance between conditions. We also include a section with unilingual versions (English and Spanish) of the critical sentences in order to determine whether bilinguals exhibit convergence so that the grammar of one of their languages permeates into the use of the other one even in unilingual mode (cf. Ebert & Koronkiewicz, 2018).

2. Motivation

This section is structured as follows. Section 2.1 discusses the AC in some detail in the context of the

theory of syntactic dependencies. Section 2.2 presents our research question and our hypotheses. Section 2.3 explains how CS can help us answer our research question.

2.1 The AC

The AC is exemplified in (1), Sentence (3) shows that placing a constituent between the verb and the direct object results in ungrammaticality in English and many other languages (see Stowell, 1981 for the original description).

(1) *Juan ate often potato chips.

(2) Juan often ate potato chips.

(3) Juan ate potato chips often.

The restriction that yields this result is extremely fine-grained: it does not affect prepositional or clausal complements:

(4) a. Mary looked carefully at him.

b. Mary said often that she didn't agree.

The contrast between (1) and (4) suggests that an obligatory syntactic dependency has failed. In (1). The adverb might intervene in the dependency or, more likely, simply signal that there is too much distance between the verb and the object. For our purposes, we do not need to figure out which grammatical principle triggers the judgments in (1-3). Instead, we ask a related question: where does the ungrammaticality of (1) originate? In contemporary syntactic theory (see Chomsky, 1995 et seq.) dependencies are based on feature checking/assignment and ungrammaticality results when a feature has

not been properly checked or assigned. Returning to (1) we can sharpen our earlier question: is this sentence ungrammatical because a feature of the verb is not satisfied or is it because a feature of the object is not satisfied?

Interestingly, the restriction is not universal. Many languages accept the equivalents of (1). The following is a Spanish example:

(5) Juan besa frecuentemente a María.
Juan kisses often to Mary
'Juan kisses Mary often.'

The root of the difference between Spanish and English has been explored in some classical work in terms of verb movement (see Pollock, 1989; Johnson, 1991 i.m.a.). This work suggests that the dependency between verb and object in Spanish is more flexible concerning its configurational requirement. Again, we do not need to enter a discussion of the linguistic principles underlying the distinction. It is enough for us to note the difference and pose it as a puzzle for linguistic theory.

2.2 Research question and hypotheses

Let's now formulate our research question. Recall that our overarching goal is to test Thurstone's method in the analysis of CS data. We have chosen the AC as our proof of concept and CS as our microscope to look at the AC. Thus, we formulate our RQ as follows:

RQ : What is the root of the AC?

As mentioned above, we assume that in a grammatical transitive sentence there is a successful syntactic dependency between the verb and the object. Moreover, this dependency is based on feature checking/assignment. But it is unclear what exactly goes wrong in an ungrammatical sentence. Is the ungrammaticality of (3a) brought about by an unchecked feature of the verb that remains unvalued at the interfaces or is it brought about by the unvalued feature of the object? Or do both need to be valued? Can these possibilities be empirically distinguished? We argue that CS data analyzed via Thurstone's method allow us to address this issue.

Four hypotheses can be considered:

H1. The unvalued Case feature of the object is responsible for the ungrammaticality of (3a). This seems to us to be the mainstream position for the last thirty years of generative grammar: a structure like (3a) violates the Case Filter (Stowell, 1981) because the object does not receive Case from a Case assigner.

H2. The unchecked feature bundle of the verb is responsible for the ungrammaticality of (3a). This solution could be developed along the lines of the Inverse Case Filter of Bošković (1997).

H3. Both the direct object and the verb contribute to the dependency and therefore unchecked features of either one may contribute to the ungrammaticality of the sentence.

H4. There is a fourth possibility that, as far as we know, has not received attention in generative grammar. The fourth possibility is that all the constituents within the domain of the dependency are somehow involved in the dependency and contribute to the acceptability or unacceptability of the structure. This

approach to sentence structure has not been tested in earnest. However, it is the one that our investigation eventually reveals as most plausible.

These hypotheses cannot be teased apart with monolingual data. An example like (3a) does not have the granularity that we need to address our RQ because we cannot extricate the contribution of the verb or the object to the (un)acceptability of a sentence. Instead, we use CS data. The type of CS that we are interested in involves the utilization of linguistic material from two lexica in the same sentence by early bilinguals who have native or native-like proficiency in both languages. We adopt the – hopefully uncontroversial - assumption that code-switching constitutes part and parcel of the rule-governed linguistic competence of such bilinguals (for detailed information on code-switching, varieties of CS and research on CS, see Bullock & Toribio, 2012).

2.3 Methodology preview: the role of code-switching

Consider the following code-switched examples (the italic part of the examples is Spanish):

(6) a. *Olivia preparó rápidamente* the food. Spanish/English
Olivia prepared quickly

b. Olivia prepared *rápidamente la comida.*

c. *Olivia preparó* quickly the food.

d. Olivia prepared quickly *la comida*.
the food

All sentences in (6) constitute classic AC configurations, with an adverb interposed between the verb and

the object. The linguistic material is drawn from a language that exhibits AC effects (English) and a language that does not (Spanish). The interesting feature of these examples is That the verb and the object appear in different languages.

Let's start with (6a,c). In these examples, the verb is in Spanish but the object is in English (for the time being, we put aside the language of the adverb). Assume (6a,c) are judged grammatical by competent bilingual speakers and code-switchers. In this circumstance, we conclude that the object is not responsible for the ungrammaticality of (3a) and therefore that it must be brought about by the verb. Correspondingly, if the sentence is judged unacceptable we conclude that the object is at least partially responsible for the AC. Likewise, if (6b,d) are grammatical, then we can conclude that the object is responsible for the AC and, if they are ungrammatical, that the verb is at least partially responsible.

Consider now the adverbs included in (6). (6a,c) and (6b,d) are different only to the extent that in one of them the adverb is in English and in the other it is in Spanish. If the adverb plays a significant role (not just as a signpost) in the acceptability of the dependency, there should be a difference between (6a) and (6c) and between (6b) and (6d).

This article reports on two experiments intended to test H1-H4. The first experiment is a traditional judgment task using Likert scales. The second experiment is formally structured around a 2AFC task and Thurstone's (1927) law of comparative judgment.

3. Experiment 1: Judgments and Likert scales

In this section we report on an experiment in which we follow the commonly used paradigm of acceptability judgments using Likert scales.

3.1 Method

3.1.1 Participants

We tested 40 early English/Spanish second-generation bilinguals; at least one of their parents was born in Mexico and all stated that they spoke the Mexican variety of Spanish. All participants stated that that they were able to speak both languages by the time they entered elementary school. A total of 27 participants stated that they learned Spanish and English simultaneously, 13 that they learned to speak Spanish first, and 2 that they learned English first. Participant demographics, as well as language characteristics, are summarized in Table 1.

Participants were recruited through Amazon Mechanical Turk, an online crowdsourcing website that has been shown a good source for acceptability judgment data (Gibson, Piantadosi, & Fedorenko, 2011), and were paid a small fee for their participation. Only workers with an acceptance rate of 95% or above and at least 100 tasks completed were allowed to take part in the study (following the guidelines proposed by Peer, Vosgerau, & Acquisti, 2014).

Before completing the main task, participants completed adapted versions of the English and Spanish Online Placement Tests used by Oxford University's Language Center (Oxford University Language Center, n.d.). The tests were modified to reflect Latin American and U.S. (rather than Spanish and U.K.) vocabulary and geographical references. Participants had to score at least 34 out of 50 points to continue with the study. This level is classified as "Higher proficiency" by the Oxford Language Center website. Overall, the pattern that emerges from the background questionnaire is that our participants were competent bilinguals, but dominant in English (see Table 1, column Experiment 1, for further details).

[Table 1 ABOUT HERE]

3.1.2 Materials

Critical trials: We generated twelve base sentences that were then modified according to four possible code-switching patterns, namely (cf. (6a-d) above):

Pattern A: Vsp+ADVsp+OBJen[1] - *Olivia preparó rápidamente* the food

Pattern B: Ven+ADVsp+OBJsp - Olivia prepared *rápidamente la comida*

Pattern C: Vsp+ADVen+OBJen - *Olivia preparó* quickly the food

Pattern D: Ven+ADVen+OBJsp - Olivia prepared quickly *la comida*

This yielded a total of 48 code-switched sentences. Each sentence consisted of 5 or 6 words and contained only one language switch (see the Appendix for a full list of the sentences used).

Filler trials: We included 44 non-critical sentences showing intra-sentential code-switching where the focus of contrast between choices was not the adverb but the determiner or the adjective. Some of the results for those trials will be reported elsewhere. By including these filler items plus the quality control items described below, critical trials made up less than a third of all items seen by participants. This was done to make it harder for raters to engage in strategic choices for their response (Cowart, 1996).

Quality control trials: There were 8 quality-control trials that consisted of sentences with inter-sentential code-switches. Each sentence had an uncontroversial error that could be easily detected if the sentences were read carefully (e.g.: *La pasé muy bien, the music* *were excellent; "I had a great time, the music were excellent"). These errors were distributed among the following factors: first vs. second half of the sentence, English vs. Spanish portion, and type of error (verb tense, number agreement, gender agreement, & word order).

[1] V = Verb, ADV = Adverb, OBJ = Object, sp = Spanish, en = English

Proper names in all sentences were chosen so that they could be used in both Spanish and English (e.g.: Max, Claudia). We avoided using nouns whose difference in onset between Spanish and English would elicit changes in the preceding indefinite article in some code-switched conditions (e.g.: "He quickly shot *a* flecha / El rápidamente disparó *an* arrow").

While there is no evidence in the literature that the type of adverb (e.g. manner, time, place, etc.) could affect our results, we decided to use only adverbs of manner (ending in *-ly* in English and *-mente* in Spanish) because they have straight-forward translation equivalents between these two languages, and because they are, by far, the most common type. Out of the 2702 English words classified as adverbs as their dominant part of speech in SubtlexUS (Brysbaert, New, & Keuleers, 2012), 83.9% are adverbs of manner (*-ly*), while in Spanish they are even more prevalent: in EsPal (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) 90.1% of the 2185 words classified as adverbs as their dominant part of speech end in *-mente*.

3.1.3 Procedure

The survey was administered online using Qualtrics in two sessions separated by about a week, with half the items presented in each session. During the first session, participants read and electronically-signed a consent form, followed by the Spanish and English tests, in that order. If they didn't achieve an acceptable score in each of those tests they were shown an "early exit" message, otherwise they were allowed to continue with the survey. At that point participants were given the choice of reading the instructions in English or Spanish. The instructions informed participants that they would see a series of sentences and that they were to indicate on a 5-point scale how "permitted" a sentence was according to the way they would speak to- or hear from another bilingual person. In the scale, a score of 1 stood for "always permitted" while 5 stood for "never permitted". Participants were then presented with the 76

code-switched sentences as described above. Each sentence was presented one at a time and the order of presentation was individually randomized for each participant. Participants had to make a choice for each item before progressing to the next one and could not go back to previous sentences. For the second session participants were given the same instructions as above, followed by the sentences to be rated. They were then given a choice of completing the background questionnaire in English or Spanish. All of the participants chose to complete the questionnaire in English.

3.2 Results

In this experiment, the independent variable was CS pattern while the dependent variable was acceptability. CS pattern had four levels, namely patterns A through D as defined in the *Materials* section. A summary of the results for this experiment can be found in Table 2.

[Table 2 ABOUT HERE]

As one of the objectives of this study is to test the reliability of our method, the analysis of the results for each session is presented independently. For the first session, we performed a one-way within-participants ANOVA that revealed a significant effect of CS pattern, $F(2.21, 86.31) = 3.63$; $p = .027$. Mauchly's test indicated that the assumption of sphericity had been violated, $\chi 2\ (5) = 27.58$, $p < .001$, therefore degrees of freedom were corrected using Greenhouse-Geisser estimates of sphericity ($\varepsilon = .74$). Post-hoc pairwise comparisons revealed that only the difference in acceptability between CS patterns B (Ven+ADVsp+OBJsp; e.g.: Olivia prepared *rápidamente la comida*) and C (Vsp+ADVen+OBJen; *Olivia preparó* quickly the food) is significant ($p = 0.03$). All other contrasts were not significant (all other p values $\geq .12$).

For the second session, we also performed a one-way within-participants ANOVA to assess the effect of CS pattern on acceptability scores. The results also show that there was a significant effect of CS pattern, $F(2.21, 86.31) = 3.63$; $p = .027$. However, post-hoc pairwise comparisons showed that none of the pairwise comparisons were significant (all p values $\geq .09$).

3.3 Preliminary Discussion

The results from Experiment 1 did not reveal a clear pattern regarding participant's acceptability of the different code-switching patterns, except maybe for a slight dis-preference for CS pattern C (Vsp+ADVen+OBJen; *Olivia preparó* quickly the food), which was only significant for the first session. We can see that the traditional Likert scale method does not provide us with any data that could help us answer our research question. First, the weak statistical power of this method has turned out data that are not granular enough: we find only one significant difference between conditions and only in the first testing session. Second, we detect a classic "race to the middle", inasmuch as results converged toward the middle of the scale. This convergence is noticeable in both sessions. To conclude, our research question remains unanswered. Given these results, we must ask whether these results were obtained because the AC is simply not present in the grammars of these subjects (even in unilingual mode) or because of limitations of the method used (in this case, most specially, lack of granularity).

In Experiment 2 we repeated the acceptability judgments for the same code-switched sentences, but this time using a two-alternative forced choice presentation of the stimulus and Thurstone's Law of Comparative Judgment for the analysis of the results. Additionally, we made sure that the AC was part of the English grammars of our participants.

4. Experiment 2: 2AFC and Thurstone's law

For ease of presentation, we introduce the two sessions under separate headings.

4.1 First session: Method

4.1.1 Participants

A total of 42 early English/Spanish second-generation bilinguals (at least one of their parents was born in Mexico and all stated that they spoke the Mexican variety of Spanish) were recruited through Amazon Mechanical Turk and were paid a small fee for completing this experiment. All participants stated that that they were able to speak both languages by the time they entered elementary school. A total of 25 participants stated that they learned Spanish and English simultaneously, 15 that they learned to speak Spanish first, and 2 that they learned English first. Their characteristics were very similar to those included in Experiment 1 and are summarized in Table 1 in the Experiment 2 column.

4.1.2 Materials

This experiment used the same materials as Experiment 1, but code-switched sentences derived from each base sentence were compared pairwise with each other in all possible combinations. The number of pairwise comparisons for each exemplar is given by the formula n*(n-1)/2, where n is the number of conditions being compared against each other. In our case, there are 4 conditions or CS patterns, so each of the 12 exemplars generates 6 pairwise comparisons for a total of 72 two-alternative forced choice critical items. The filler sentences from Experiment 1 were also presented as 2AFC items (156 comparisons), and there were an additional 8 quality control sentences for a total of 24 comparisons for a total of 252 items on the survey. The main block for each testing session thus consisted of 126 comparisons (half of each type described above).

Unilingual sentences: In a separate block of the survey, participants were presented with English and

Spanish unilingual versions of the base sentences used in the main part of the survey. They had to choose between versions of each sentence where the adverb was presented pre- or post-nominally; in other words, participants had to choose between Verb+Object constructions and Verb+Adverb+Object constructions. English and Spanish sentences were presented separately. We included these unilingual sentences to assess for possible grammar convergence between languages in unilingual mode (see the Appendix for a full list of sentences).

4.1.3 Procedure

The procedure was very similar to that of Experiment 1, but the instructions informed participants that they would see a series of sentence pairs, and asked them to pick the one closer to the way they would speak to another bilingual person, asking to make a choice even if both sentences sounded "right" or both sounded "wrong" (as opposed to rating each individual sentence on a 5-point scale as in Experiment 1). The pairs of sentences were presented one at a time and the order of presentation of the pairs, as well as the order of each sentence within each pair, was individually randomized for each participant. After the main block, they would see the Spanish or the English unilingual blocks. The order of these two blocks was also randomized for each participant. Finally, they were given a choice of completing the background questionnaire in English or Spanish. All but one participant chose to complete the questionnaire in English.

4.2. Results of session 1

4.2.1 Unilingual Sentences

For unilingual English sentences, participants chose Verb+Object constructions (E.g.: Olivia prepared the food quickly; No violation of Adjacency Condition) in 98% of the trials. For unilingual Spanish

sentences, as expected, there was some variation: participants chose Verb + Object constructions (*Olivia preparó la comida rápidamente*) in 72.2% of the trials and Verb + Adverb + Object (*Olivia preparó rápidamente la comida*) in the rest of the trials. Taken in tandem, these results indicate that the adjacency condition is clearly a feature of the English grammar of these English/Spanish bilinguals and that it is not a feature that is vulnerable to a contact/immersion situation. The fact that Spanish speakers allowed the adverb to stand between the verb and the object about 28% of the time indicates that for these speakers there is no categorical restriction to placing an adverb between the verb and the object - no AC, in other words. Speakers tend to prefer the V+O+Adv order because this is the neutral order in out-of-the-blue sentences. The order V+Adv+O favors narrow focus on the object, which requires a particular type of context to be fully felicitous in a discourse (see Zubizarreta 1998 for the classical description).

4.2.2 Code Switched-Sentences

Participants' responses for the main block of trials were analyzed using Thurstone's (1927) Law of Comparative Judgment, Case V.

Table 3 shows the rank order and "measure" for each condition. The measure values are relative to the pattern with the lowest acceptability (which is by convention set to 0). As explained in the introduction, the unit of measure on the scale is the standard deviation of the data and each of those values represents the center of the discriminal distribution for each condition. The 95% confidence interval for this set of data was 0.06. A within-participants ANOVA revealed a significant effect of sentence type $F(3,1004) = 26.81$, $p < .001$. Post-Hoc Tukey HSD tests indicated that all contrasts were highly significant (all p values < .001) except for the contrast between sentence patterns D (Ven+ADVen+OBJsp; e.g.: Olivia prepared quickly *la comida*) and B (Ven+ADVsp+OBJsp; e.g.: Olivia

prepared *rápidamente la comida*) (p = .99). Overall, there was a clear preference for pattern A (Vsp+ADVsp+OBJen; e.g.: *Olivia preparó rápidamente* the food) over all other patterns.

It is important to highlight that the 0 score for condition C (Vsp+ADVen+OBJen; e.g.: *Olivia preparó* quickly the food) does not indicate that no participant chose that pattern, but rather that as the least favorite option it is chosen as a point of comparison for the other options (an arbitrary origin or "0-point" being one of the characteristics of interval scales).

[Table 3 ABOUT HERE]

In the following section we include a formal description of how the numbers in this table were obtained.

4.2.3 Calculation of the Thurstone measure

The formal aspects of Thurstone's analysis are discussed in great detail in various sources (David, 1988; Green & Swets, 1966; Thurstone, 1927; Torgenson, 1958), and although the statistical formulas presented there may seem a bit daunting, the essence of the method is fairly straightforward. Below is a description, in layman's terms, of the steps needed to calculate Thurstone's measure using a simple spreadsheet illustrated with examples from our calculations for the results of Experiment 2, Session 1 (Please refer to Figure 1 for the examples):

[Figure 1 ABOUT HERE]

Step 1: Determine, for each comparison (A vs B, B vs C, etc.), the number of times each option was chosen when contrasted with each of the other options and arrange them into a matrix (Figure

1a).

Step 2: From those values, calculate the *proportion* of times each option was a winner or loser against all other options by dividing them by the total number of data points for each comparison. In our case, we had 6 exemplars per comparison and 42 participants, so we divide each entry in the matrix by 252 (Figure 1b).

Step 3: Transform each entry in the matrix to a Z score (Tip: in order to calculate the Z scores we used the Excel formula NORM.S.INV, "Inverse of the standard normal cumulative distribution") (Figure 1c).

Step 4: Multiply each of those Z scores by the square root of 2 (Figure 1d).

Step 5: Take the average of each row in the matrix. For example, for A we would average 0.3120, 0.6091, and 0.2832 to yield 0.401 (Figure 1e).

Step 6: We now apply a linear transformation to those scores so that all are positive numbers. We do this by finding the smallest score, in this case -.397 and adding it to each of the scores. This shifts the origin for all values, in effect making the lowest score the point of comparison for all other scores.

Step 7: Those values are the Thurstone scores for each of our options. We now just need to rank them in descending order to find their relative position in an interval scale, that is, on in which the distance (though not the ratio) between its values is meaningful. The values thus obtained can then be tested using standard statistical methods such as standard errors and ANOVA.

While the use of scales in the collection of acceptability judgments is widespread, there have been some concerns regarding their stability and reliability, both between informants (e.g.: Bader & Häussler, 2010; Labov, 1972, 1975; Ross, 1979; Stokes, 1974) and for the same informants on different occasions

(Caroll, Bever, & Pollack, 1981; Nagata, 1988; Snow & Meijer, 1977). In order to test the reliability and stability of our results, we asked participants to complete a second 2AFC survey about a week after the first one with different instances of the same grammatical constructions tested before.

4.3 Second session: Method

4.3.1 Participants, Materials, and Procedure

Of the original 42 participants that completed the first session, 36 returned for this second session. The procedure was identical to that of the first session, except that participants did not have to complete the language proficiency tests and the background questionnaire. The six base sentences used for this second part, as well as the filler and quality control items, had the same structures as those described for Experiment 1, but the actual sentences were all different (See the Appendix for a list of base sentences).

4.3.2 Results of session 2

Table 2 shows the results for this experiment. The 95% confidence interval for this set of data was 0.07. A within-participants ANOVA revealed a significant effect of sentence type $F(3, 864) = 23.48$, $p < .001$. In terms of the individual comparisons, the Post-Hoc Tukey HSD tests showed a pattern very similar to the first experiment (most contrasts highly significant with p values < .001, with the exception of the contrast between patterns A and D with $p < .05$). Once again there was no significant difference between patterns D and B ($p = .48$).

For unilingual comparisons, in English participants preferred the Verb+Object construction over the Verb+Adverb+Object construction in 96.6% of the trials, while in Spanish they chose the Verb+Object construction in 67.6% of the trials.

As can be seen, the results for the second session were highly consistent with those of the first

session, attesting to the within-rater reliability of our results. Since the same pattern of results was obtained with a different set of items, we can also be assured of the generalizability of our results to this type of construction. Our method provides very stable results.

5. Discussion

As mentioned, the results from the unilingual sentences show that the AC is a feature of the English grammar of the bilingual Spanish/English speakers tested. Correspondingly, the AC is absent from the Spanish grammar of the same speakers as a categorical restriction - instead, what we see is a preference for a certain word order in a neutral context, a restriction based on pragmatics and not grammar. Thus, we have a solid springboard from which to analyze the code-switched sentences.

Our results show that the 2AFC and Thurstone's Law have proven to be an excellent method to extract grammaticality judgments. The judgments do not vary substantially among comparable subjects; they can be replicated in different sessions with a time interval in between them; most importantly, they present a granularity that was absent in the Likert scale experiment; granular enough, in fact, to test our research question, as we show below. We must conclude that the acceptability cline that our experiments show and is reflected in Table 3 is not random but a reflection of the linguistic competence of the participating subjects, puzzling as it may seem.

Let's see if our results support any of the hypotheses laid out in section 1. H1 is not supported, H1 would predict that all sentences that include a Spanish verb should be preferred. But that is not the case – in fact, pattern C (Vsp+ADVen+OBJen; e.g.: *Olivia preparó* quickly the food), which includes a Spanish verb, was the least acceptable pattern. Likewise, H2 is not supported either. H2 predicts that all sentences that include a Spanish object should be preferred. But pattern A (Vsp+ADVsp+OBJen; e.g.: *Olivia preparó rápidamente* the food) is our subjects' favorite, and it includes an English object. H3 should also

be rejected because of the existence of the cline itself. H3 suggests that the AC is a combination of the features of *v* and the object. H3 does not predict a cline, all patterns include a combination of *v* and object in different languages and therefore H3 predicts there should be no significant differences between the patterns. The only hypothesis consistent with the result is H4: all the constituents within the domain of the dependency contribute to the perceived (un)acceptability of the construction.

Let's now look at the data in more detail. There are two observations that we can extract out of the data. The first one is that patterns A, B, D, have at least the adverb or the object in Spanish. The lowest ranked pattern, C (Vsp+ADVen+OBJen; e.g.: *Olivia preparó* quickly the food), has both the adverb and the object in English. The second observation is that pattern A has the verb in Spanish while patterns B (Ven+ADVsp+OBJsp; e.g.: Olivia prepared *rápidamente la comida*) and D (Ven+ADVen+OBJsp; e.g.: Olivia prepared quickly *la comida*) have the verb in English. Thus, the generalization that emerges from these results is that if at least the adverb or the object is in Spanish, having the verb also in Spanish improves the acceptability of the sentence. But if neither the adverb nor the object are in Spanish, the Spanish verb does not improve the sentence.

An anonymous reviewer suggests that we look at it from a different perspective. Noticing that patterns A and D do not involve a switch between the verb and the adverb while patterns B and C do, the reviewer proposes that this fact should be incorporated into the analysis and conclude that what induces unacceptability is CS between verb and adverb. However, we do not agree with this conclusion because the acceptability distance between D and B did not turn out to be significant in either session. Therefore we feel we do not have solid grounds for the hypothesis that switching between verb and adverb is dispreferred. Moreover, if we had in fact found that the difference between B and D is significant, this would not shed any light on why A is so clearly preferred to D and B to C.

How should we interpret these results? We would like to suggest a tentative analysis, in need of much

deeper work to explore its consequences. Assume a structure for transitive predicates as in (10) (see Chomsky, 1995 et seq., Kratzer, 1996):

(7) EA *v* [$_{VP}$ Adv [$_{V'}$ V IA]]

(7) encodes a series of assumptions. The first one is that a lexical verb merges with the internal argument to form a verbal projection. An adverb can be merged as an adjunct to the verbal phrase. The VP is the complement of a higher predicate, called *v* or "little *v*", which encodes the type of event semantics. If the event is of the right type, *v* introduces an external argument.

The lexical verb and *v* may remain as separate syntactic terminals in a type of light verb construction. Most often, the lexical verb raises and incorporates into *v*:

(8) (EA) *v*+V [$_{VP}$ Adv [$_{V'}$ t(V) Obj]]

Further, assume a *phase*-based theory of locality, according to which the complement of *v transfers* as a unit to the interpretive systems, the intentional-conceptual system and the externalization system (Chomsky 2000). In (8), this entails that the VP constituents – the adverb and the internal argument – transfer in one shot.

As an anonymous reviewer points out, the assumption that the representation in (8) underlies the structure of all English and Spanish predicates is simplistic. There is good evidence that at least some English objects undergo short scrambling (Johnson, 1991) and the Differential Object Marking phenomena in Spanish points in the same direction (López, 2012). For us, what is important is that the object and the adverb are located within a constituent that excludes the verb.

Within these assumptions, we claim the following: the AC is an interface condition that constrains an interface representation as it transfers to the externalization systems (we find it unlikely that the AC has anything to do with semantic interpretation). In particular, the AC is triggered by an English-lexicon item when the VP is transferred to the interpretive systems.[2] The English-lexicon item can be the adverb or the object. An AC violation is attenuated if there is at least one VP constituent in Spanish. Additionally, if at least one of the VP constituents is in Spanish, having the V+v in Spanish further raises the acceptability of the sentence – probably as a reflex of having a copy of the Spanish V in the VP. If no constituent within the VP is in Spanish the result is a sharp drop in acceptability. The study of the implications of these findings lies outside of the scope of the present study.

6. Conclusions

In this paper, we have tried out two methods to obtain acceptability judgments in CS: judgments on a Likert scale and a 2AFC task combined with a Thurstone measurement model. We have shown that the 2AFC is superior to the traditional Likert scale because it yields more granular data and this allowed us to approach our research question. Additionally, we have shown that the 2AFC task provided us with reliable results. We believe that the results of our experiments should lay to rest the notion that judgments on CS sentences are not possible - let alone the profoundly mistaken notion that CS is somehow not an expression of a person's linguistic competence. More generally, we surmise that 2AFC is a suitable method to study linguistic competence. We also offer a step-by-step explanation of how to implement the analysis that should facilitate its adoption by other researchers in the field of code-switching (and

[2] The formulation of the restriction in this paragraph would suggest that we subscribe to the notion that bilinguals have two separate lexicons that get mingled in CS (as in MacSwan 1999 i.m.a.) In fact, for the purposes of this article, we remain agnostic with respect to this assumption. Where it says "English(Spanish)-lexicon" we refer to some property present in items that spell-out as words that we would recognize as "English(Spanish)". At this point, our understanding of the issues does not allow us to delve deeper.

linguistics in general). Additionally, the clarity of the results obtained in this study should serve as encouragement to continue exploring how some of the very sophisticated methods that have been developed to measure judgement data in the field of psychophysics can be applied to the field of acceptability judgments in linguistic research.

Our research question has received an unexpected answer: it turns out that the AC cannot be traced to any one constituent but it is a property that emerges when several constituents appear together in the same structure. This is a conclusion with abundant ramifications that we leave for future research.

REFERENCES

Anderson, T. (2006). Spanish-English bilinguals' attitudes toward code-switching: proficiency, grammaticality, and familiarity. Doctoral dissertation, The Pennsylvania State University, State College, Pennsylvania.

Bader, M., & Häussler, J. (2010) Toward a model of grammaticality judgments. *Journal of Linguistics,* 46, 273-330.

Badiola, L., Delgado, R., Sande, A., & Stefanich, S. (2018) Code-switching attitudes and their effects on acceptability judgment tasks. *Linguistic Approaches to Bilingualism*, 8(1). DOI: 10.1075/lab.16006.bad

Bader, M., & Häussler, J. (2010) Toward a model of grammatical judgments. *Journal of Linguistics, 46*, 273-330.

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

Bialystok, E. (1979). Explicit and implicit judgements of L2 grammaticality. *Language Learning, 29*, 81-103.

Bock, R. D., & Jones, L.V. (1968) *The measurement and prediction of judgment and choice*. San Francisco, CA: Holden-Day.

Bošković, Ž. (1997) *The Syntax of Nonfinite Complementation: An Economy Approach*. Cambridge, MA: MIT Press.

Brown, T.C., Peterson, G.L. (2009) *An enquiry into the method of paired comparison: reliability, scaling,*

and Thurstone's Law of Comparative Judgment. Fort Collins, CO: U.S. Department of Agriculture, Forest Service.

Brysbaert, M., New, B., & Keuleers, E. (2012) Adding Part of Speech information to the SUBTLEXUS word frequencies. *Behavior Research Methods, 44*, 991-997.

Bullock, B.E. & Toribio, A.J. (2012) *The Cambridge Handbook of Linguistic Code-switching.* Cambridge, UK: Cambridge University Press.

Caroll, J.M., Bever, T.G., & Pollack, C.R. (1981) The non-uniqueness of linguistic intuitions. *Language, 57*, 368-383.

Cattelan, M. (2012) Models for paired comparison data: a review with emphasis on dependent data, *Statistical Science, 27*, 412–433

Chomsky, N. (1957) *Syntactic Structures.* The Hague, the Netherlands: Mouton.

Chomsky, N. (1986) *Knowledge of Language: Its nature, origin and use*. New York: Praeger.

Chomsky, N. (1995) *The Minimalist Program*, Cambridge, MA: MIT Press.

Chomsky, N. (2000) Minimalist Inquiries: The Framework. In R. Martin, D. Michaels & J. Uriagereka (Eds.) *Step by Step: Essays in Minimalist Syntax in Honor of Howard Lasnik*, 89-155. Cambridge, MA: MIT Press.

Cowart, W. (1996) *Experimental Syntax: Applying Objective Methods to Sentence Judgments*. California: Sage Publications Inc.

David, H.A. (1988) *The method of paired comparisons* (2nd ed.). New York, NY: Oxford University Press.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013) EsPal: One-stop shopping for Spanish word properties, *Behavior Research Methods, 45*, 1246–1258

Ebert, S. & Koronkiewicz, B. (2017) Monolingual stimuli as a foundation for analyzing code-switching data, *Linguistic Approaches to Bilingualism*, *8*(1).

Edwards, L. (1957) *Techniques of Attitude Scale Construction.*Englewood Cliffs, New Jersey: Prentice-Hall.

Engen, T. (1971) Psychophysics, Vol. II: Scaling methods. In J. Kling & L. Riggs (Eds.) *Woodworth and Schlosberg's experimental psychology*, 89-91. New York: Holt, Rinehart, & Winston.

Featherston, S. (2005a) Magnitude estimation and what it can do for your syntax: Some WH-constructions in German. *Lingua, 115*, 1525-50.

Featherston, S. (2005b). Universals and grammaticality: wh-constraints in German and English. *Linguistics*, *43*, 667–711.

Featherson, S. (2008) Thermometer judgments as linguistic evidence. In C. M. Riehl and A. Rothe, eds. *Was ist linguistische Evidenz?* Aachen: Shaker Verlag, 69–89.

Fukuda, Shin, Grant Goodall, Dan Michel and Henry Beecher (2012) Is Magnitude Estimation worth the trouble? In Jaehoon Choi, E. Alan Hogue, Jeffrey Punske, Deniz Tat, Jessamyn Schertz and Alex Truman (eds.), *Proceedings of the 29th West Coast Conference on Formal Linguistics* (WCCFL29). 328-336. Somerville, MA: Cascadilla Proceedings Project.

Giancaspro, D. (2013) L2 Learners' and Heritage Speakers' Judgments of Code-Switching at the Auxiliary-VP Boundary. Selected Proceedings of the 16th Hispanic Linguistics Symposium, ed. Jennifer Cabrelli Amaro et al., 56-69. Somerville, MA: Cascadilla Proceedings Project.

Gibson, E., Piantadosi, S., & Fedorenko, K. (2011) Using Mechanical Turk to Obtain and Analyze English Acceptability Judgments. *Language and Linguistics Compass* 5, 509-524.

Gigerenzer, G., & Richter, H. (1990) Context effects and their interaction with development: Area judgments. *Cognitive Development, 5*, 235–264.

Gigerenzer, G., Krauss, S. & Vitouch, O. (2004) The null ritual: What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (ed.) *The Sage handbook of quantitative methodology for the social sciences*. Thousand Oaks, CA: Sage.

Green, D.M. & Swets J.A. (1966) *Signal Detection Theory and Psychophysics*, Wiley, New York.

Guilford, J. P. (1954) *Psychometric methods* (2d ed.). New York,: McGraw-Hill.

Haegeman, L. (1991) *Introduction to government and binding theory*. Oxford: Blackwell.

Johnson, K. (1991) Object Positions. *Natural Language and Linguistic Theory, 9*, 577–637.

Jones, L.V., & Thissen, D. A. (2007) A History and Overview of Psychometrics. In C.R. Rao & S. Sinharay (eds.) *Handbook of Statistics, volume 26: Psychometrics*, 1-27. New York, NY: Elsevier.

Kratzer, A. (1996) Severing the External Argument from its Verb, in J. Rooryck & L. Zaring (eds.): *Phrase Structure and the Lexicon*. Dordrecht: Kluwer Academic Publishers.

Labov, W. (1972) Some principles of linguistic methodology, *Language in Society* 1, 97-120.

Labov, W. (1975) *What is a linguistic fact?* Lisse: Peter de Ridder.

López, L. (2012) *Indefinite objects: scrambling, choice functions and differential marking*. Cambridge, Mass: MIT Press.

Luce, R. D. (2002) A psychophysical theory of intensity proportions, joint presentations, and matches. *Psychological Review, 109*. 520-532.

MacSwan, J. (1999) *A minimalist approach to intrasentential code switching: Spanish-Nahuatl bilingualism in Central Mexico*. New York: Garland.

Montag, E.D. (2006) Empirical formula for creating error bars for the method of paired comparisons. *Journal of Electronic Imaging*, *15*, 222-230.

Mohan, B.A. (1977) Acceptability testing and fuzzy grammar. In Sidney Greenbaum (ed.), *Acceptability*

in language, 133–148. The Hague: Mouton

Muysken, P. (2000) *Bilingual Speech: A typology of code-mixing*. Cambridge: Cambridge University Press

Munarriz, A. & Parafita Couto, M.C. (2014) ¿ Cómo estudiar el cambio de código ? Incorporación de diferentes metodologías en el caso de varias comunidades bilingües. *Lapurdum, 18*, 43-73

Nagata, H. (1988) The relativity of linguistic intuition: The effect of repetition on grammaticality judgments. *Journal of Psycholinguistic Research*, 171, 1-17.

Narens, L. (1996) A theory of ratio magnitude estimation. Journal of Mathematical Psychology, 40, 109-129

Nunnally, J.C. (1967) *Psychometric theory*. New York: McGraw Hill.

Onar Valk, P. (2014) Convergent developments in Dutch Tukish word order - A comparative study using 'elicited production' and 'judgment' data: Converging evidence?, *Applied Linguistics Review, 5*, 353-374.

Oxford University Language Centre. "Placement Tests." lang.ox.ac.uk. http://www.lang.ox.ac.uk/tests/index.html (accessed July 1st, 2015)

Parafita Couto M.C., Deuchar M., & Fusser, M. (2015) How do Welsh-English bilinguals deal with conflict? Adjective-noun order resolution. In: G. Stell, K. Yakpo (Eds.) *Code-switching at the crossroads between structural and sociolinguistic perspectives*, 65-84. Boston: Mouton de Gruyter.

Parafita Couto, M. C., Munarriz, A., Epelde, I., Deuchar, M., & Oyharçabal, B. (2015) Gender conflict resolution in Spanish-Basque mixed DPs. *Bilingualism, Language and Cognition, 18*, 304–323.

Parraga, C.A. (2015) Perceptual Psychophysics. In G. Cristobal, M. Keil, & L. Perrinet (Eds.) *Biologically-Inspired Computer Vision: Fundamentals and Applications*, 81-108. New York, NY:

Wiley.

Peer, E., Vosgerau, J., & Acquisti, A. (2014) Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*, 1023-1031

Pollock, J.Y. (1989) Verb Movement, Universal Grammar, and the Structure of IP. *Linguistic Inquiry, 20*, 365-424

Reber, A. (1995) *The Penguin dictionary of psychology*. New York: Penguin Books.

Ross, J.R. (1979) Where's English? In C.J. Fillmore, D. Kemper, & W.S. Wang (Eds.) *Individual differences in language ability and language behavior*, 127-163. New York: Academic Press.

Serratrice, L., Sorace, A., Filiaci, F., & Baldo, M. (2009) Bilingual children's sensitivity to specificity and genericity: Evidence from metalinguistic awareness. *Bilingualism: Language and Cognition, 12*, 239-257.

Schütze, C.T. (2016) *The empirical base of linguistics. Grammaticality judgments and linguistic methodology*. Chicago & London: University of Chicago Press.

Schütze, C.T., & Sprouse, J. (2013) Judgment Data. In R.J. Podesva and D. Sharma (eds.), *Research Methods in Linguistics*, pp. 27-50. Cambridge, England: Cambridge University Press.

Snow, C., & Meijer, G. (1977) On the secondary nature of syntactic intuitions. In S. Greenbaum (Ed.) *Acceptability in language*, 163-177. The Hague, the Netherlands: Mouton.

Sorace, A. (1996) The use of acceptability judgments in second language acquisition research. In W. C. Ritchie and T. K. Bhatia (eds.), *Handbook of second language acquisition*, pp. 375–409. San Diego, CA: Academic Press.

Sorace, A. (2010) Magnitude estimation in language acquisition research. In S. Unsworth and E. Blom (Eds.), *Experimental Methods in Language Acquisition*, pp. 57-72. Amsterdam: John Benjamins.

Sorace, A., & Keller, F. (2004) Gradience in linguistic data. Lingua, 115, 1497-1524.

Sorace, A. Serratrice, L., Filiaci, F., & Baldo, M. (2009) Discourse conditions on subject pronoun realization: Testing the linguistic intuitions of older bilingual children. *Lingua, 119*, 460-477.

Sprouse, J. (2007) Continuous acceptability, categorical grammaticality, and experimental syntax. *Biolinguistics*, 1, 118-129.

Sprouse, J. (2011) A Test of the Cognitive Assumptions of Magnitude Estimation: Commutativity does not Hold for Acceptability Judgments. *Language, 87*, 274-288.

Sprouse, J. & Almeida, D. (2011) Power in acceptability judgment experiments and the reliability of data in syntax. Ms., University of California, Irvine & Michigan State University.

Sprouse, J. & Almeida, D. (2012) The empirical status of data in syntax: A reply to Gibson and Fedorenko. *Language and Cognitive Processes*. iFirst. 1–7.

Sprouse, J., Schütze, C.T. & Almeida, D. (2013) A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001-2010. *Lingua, 134*, 219–248.

Stevens, S. S. (1946) On the Theory of Scales of Measurement. *Science*, 103, 677–680.

Stevens, S.S. (1956) The direct estimation of sensory magnitudes: loudness. *The American journal of psychology, 69*, 1–25

Stowell, T. (1981) Origins of phrase structure. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Stokes, W. (1974) All of the work on quantifier-negation isn't convincing. In M.W. La Galy, R.A. Fox, & A. Bruck (Eds.) *Papers from the tenth regional meeting, Chicago Linguistic Society*, 692-700. Chicago: Chicago Linguistic Society.

Thurstone, L. (1927) A Law of Comparative Judgment. *Psychological Review, 34*, 273-286.

Tikofsky, R.S., & Reiff, D.G. (1970) Structural apperception in the absence of syntactic constraints, *Language and Speech, 13*, 240-253.

Tsukida, K. & Gupta, M.R. (2011) *How to Analyze Paired Comparison Data* (UWEE Technical Report Number UWEETR-2011-0004) Seattle, University of Washington.

Torgerson, W.S. (1958) *Theory and methods of scaling*. New York, NY: John Wiley & Sons

Toribio, A.J. (2001) On the emergence of bilingual code-switching competence. *Bilingualism: Language and Cognition*, 4, 203–231.

Weskott, T. & Fanselow, G. (2008) Variance and informativity in different measures of linguistic acceptability. *Proceedings of the West Coast Conference on Formal Linguistics* (WCCFL), *27*, 431-439.

Weskott, T., & Fanselow, G. (2011) On the Informativity of Different Measures of Linguistic Acceptability. *Language, 87*, 249–273

Table 1.

Participant characteristics for experiments 1 and 2.

	Experiment 1	Experiment 2
Total number of participants	40	42*
Number of Female/Male participants	18/22	23/19
Number of participants born in the U.S.	36	35
Mean age of immigration to U.S. in years (for those not born in U.S.)	4:6 (SD: 1.7)	4:5 (SD: 1.8)
English exam score (out of 50)	44.6 (SD: 2.8)	45.2 (SD: 2.7)
Spanish exam score (out of 50)	42.9 (SD: 4.3)	41.9 (SD: 4.5)
Self-assessment of English proficiency**	3.98 (SD: 0.2)	3.95 (SD: 0.2)
Self-assessment of Spanish proficiency**	3.60 (SD: 0.6)	3.40 (SD: 0.7)
Participants with Spanish-only maternal input while growing up	21	27
Participants with Spanish-only paternal input while growing up	17	23
Participants with English-only elementary school	29	29
Participants with English-only high school	33	38
Participants that speak mostly English in their workplace	31	33
Participants that speak both English and Spanish with friends and family	20	24
State of residence	California: 20 Texas: 14 New Jersey: 3 Arizona: 2 Colorado: 1	California: 29 Texas: 12 Arizona: 1

*Only 36 participants returned to do the second session of Experiment 2
** On a scale of 1 to 4, where 4 indicates “Confident in extended conversations”

Table 2.

Summary of results for Experiment 1.

Pattern	Construction	Example	Mean Acceptability Score	
			Session 1	Session 2
A	Vsp+ADVsp+OBJen	*Olivia preparó rápidamente* the food	2.68 (SD:1.00)	2.65 (SD: 0.97)
B	Ven+ADVsp+OBJsp	Olivia prepared *rápidamente la comida*	2.67 (SD:0.89)	2.63 (SD: 1.03)
C	Vsp+ADVen+OBJen	*Olivia preparó* quickly the food	2.99 (SD:1.06)	2.83 (SD: 1.04)
D	Ven+ADVen+OBJsp	Olivia prepared quickly *la comida*	2.82 (SD:0.99)	2.58 (SD: 1.02)

Table 3.

Ranking and measure for sentences presented in Experiment 2

				Thurstone Measure	
Rank	Pattern	Construction	Example	Session 1	Session 2
1	A	Vsp+ADVsp+OBJen	*Olivia preparó rápidamente* the food	0.80	0.79
2	D	Ven+ADVen+OBJsp	Olivia prepared quickly *la comida*	0.41	0.54
3	B	Ven+ADVsp+OBJsp	Olivia prepared *rápidamente la comida*	0.38	0.44
4	C	Vsp+ADVen+OBJen	*Olivia preparó* quickly the food	0.00	0.00

	Loser			
Winner	A	B	C	D
A	-	148	168	146
B	104	-	140	131
C	84	112	-	99
D	106	121	153	-

(a) Step 2

	Loser			
Winner	A	B	C	D
A	-	0.59	0.67	0.58
B	0.41	-	0.56	0.52
C	0.33	0.44	-	0.39
D	0.42	0.48	0.61	-

(b) Step 2

	Loser			

Winner	A	B	C	D
A	-	0.221	0.431	0.200
B	-0.221	-	0.140	0.050
C	-0.431	-0.140	-	-0.272
D	-0.200	-0.050	0.272	-

(c) Step 3

	Loser			
Winner	A	B	C	D
A	-	0.312	0.609	0.283
B	-0.312	-	0.198	0.070
C	-0.609	-0.198	-	-0.384
D	-0.283	-0.070	0.384	-

(d) Step 4

	Average
A	0.799
B	0.382
C	0.000
D	0.407

(f) Step 6

	Average
A	0.401
B	-0.015
C	-0.397
D	0.010

(e) Step 5

Figure 1. Illustration of the steps for Thurstone's Analysis

Appendix

Sentences used in Experiments 1 and 2:

Monolingual base sentences:

Block 1

English	Spanish
Olivia prepared quickly the food	*Oliva preparó rápidamente la comida*
Alan answered wisely the questions	*Alan respondió sabiamente las preguntas*
Claudia took politely the spoon	*Claudia tomó educadamente la cuchara*
Victor cleans frequently the house	*Victor limpia frecuentemente la casa*
Gabriel confronted bravely the problem	*Gabriel confrontó valientemente el problema*
The player kicked crazily the ball	*El jugador pateó locamente la pelota*

Block 2

English	Spanish
Max watched carefully the demonstration	*Max observó cuidadosamente la demostración*
David broke accidentally the cup	*David quebró accidentalmente la taza*
Sonia obeyed silently the order	*Sonia obedeció silenciosamente la orden*
The student read nervously the message	*El estudiante leyó nerviosamente el mensaje*
Lucas kissed tenderly the picture	*Lucas besó tiernamente la foto*
Clara closed firmly the door	*Clara cerró firmemente la puerta*

Code switched sentences used in Experiments 1 and 2:

Condition A: Vsp+ADVsp+OBJen

Block 1

Olivia preparó rápidamente the food

Alan respondió sabiamente the questions

Claudia tomó educadamente the spoon

Victor limpia frecuentemente the house

Gabriel confrontó valientemente the problem

El jugador pateó locamente the ball

Block 2

Max observó cuidadosamente the demonstration

David quebró accidentalmente the cup

Sonia obedeció silenciosamente the order

El estudiante leyó nerviosamente the message

Lucas besó tiernamente the picture

Clara cerró firmemente the door

Condition B: Ven+ADVsp+OBJsp

Block 1

Olivia prepared *rápidamente la comida*

Alan answered *sabiamente las preguntas*

Claudia took *educadamente la cuchara*

Victor cleans *frecuentemente la casa*

Gabriel confronted *valientemente el problema*

The player kicked *locamente la pelota*

Block 2

Max watched *cuidadosamente la demostración*

David broke *accidentalmente la taza*

Sonia obeyed *silenciosamente la orden*

The student read *nerviosamente el mensaje*

Lucas kissed *tiernamente la foto*

Clara closed *firmemente la puerta*

Condition C: Vsp+ADVen+OBJen

Block 1

Olivia preparó quickly the food

Alan respondió wisely the questions

Claudia tomó politely the spoon

Victor limpia frequently the house

Gabriel confrontó bravely the problem

El jugador pateó crazily the ball

Block 2

Max observó carefully the demonstration

David quebró accidentally the cup

Sonia obedeció silently the order

El estudiante leyó nervously the message

Lucas besó tenderly the picture

Clara cerró firmly the door

Condition D: Ven+ADVen+OBJsp

Olivia prepared quickly *la comida*

Alan answered wisely *las preguntas*

Claudia took politely *la cuchara*

Victor cleans frequently *la casa*

Gabriel confronted bravely *el problema*

The player kicked crazily *la pelota*

Block 2

Max observó carefully the demonstration

David quebró accidentally the cup

Sonia obedeció silently the order

El estudiante leyó nervously the message

Lucas besó tenderly the picture

Clara cerró firmly the door