2023

Measuring bilingual language dominance: An examination of the reliability of the Bilingual Language Profile

Daniel J. Olson
Purdue University, danielolson@purdue.edu

Recommended Citation

Olson, D. J. (2023). Measuring bilingual language dominance: An examination of the reliability of the Bilingual Language Profile. *Language Testing,* 1–45. https://doi.org/10.1177/02655322221139162

Measuring bilingual language dominance: An examination of the reliability of the Bilingual * Language Profile

Daniel J. Olson

Purdue University
640 Oval Dr., West Lafayette, IN, USA, 47907
danielolson@purdue.edu

Abstract:
Measuring language dominance, broadly defined as the relative strength of each of a bilingual's two languages, remains crucial methodological issue in bilingualism research. While various methods have been proposed, the Bilingual Language Profile (Birdsong et al., 2012) has been one of the most widely used tools for measuring language dominance. While previous studies have begun to establish its validity, the Bilingual Language Profile has yet to be systematically evaluated with respect to reliability. Addressing this methodological gap, the current study examines the reliability of the Bilingual Language Profile, employing a test-retest methodology with a large ($N = 248$), varied sample of Spanish–English bilinguals. Analysis focuses on the test-retest reliability of the overall dominance score, the dominant and non-dominant global language scores, and the subcomponent scores. Results demonstrate that language dominance score produced by the BLP shows 'excellent' levels of test-retest reliability. In addition, while some differences were found between the reliability of global language scores for the dominant and non-dominant languages, and for the different subcomponent scores, all components of the BLP display strong reliability. Taken as a whole, this study provides evidence for the reliability of Bilingual Language Profile as measure of bilingual language dominance.

Keywords: bilingualism, language dominance, proficiency, reliability, questionnaire

1. Introduction

In research on bilingual populations, among the most basic and long-standing questions has been how to assess a bilingual's abilities in each of their two languages. Although the general public may view a bilingual as someone who has equal "mastery" in each of their two languages (Wei, 2000, p. 6), the field of bilingualism has taken a much broader approach, generally describing a bilingual as someone with knowledge of two languages, or who uses two languages or language varieties in everyday interactions (e.g., Grosjean, 2008; Montrul, 2016). This broad conceptualization of bilingualism encompasses both those who use their two languages across a variety of interactional contexts and those who are in the process of actively acquiring a second language (L2). Given the broad spectrum of speakers considered to be bilingual, the issue of measuring a bilingual's language abilities has been of importance in both describing the linguistic profile of bilinguals and accounting for variation in bilingual behaviors from both research and clinical perspectives (e.g., Gertken et al., 2014). In assessing a bilingual's abilities, two key components are considered – a bilingual's ability in each of their languages individually (i.e., proficiency) and the *relative* strength of the two languages (i.e., dominance).

In developing useful assessments for the field, researchers should strive to create methods that are both valid and reliable. Validity has been variously defined, with somewhat different perspectives found in psychometric and language testing literature. From a psychometric perspective, validity concerns whether the assessment method reflects the underlying construct and correlates with other subjective and objective assessments of the same construct (Dörnyei & Taguchi, 2009). From an educational or language testing perspective (for discussion of an argument-based approach see Chapelle, 2011; Kane, 2016), validity is contextualized within the

proposed interpretations and uses of the measure, and evaluated with respect to the "appropriateness" of the measure for a given purpose (Kane, 2016, p. 198). With respect to reliability, the assessment should evidence a high degree of measurement stability or a low degree of measurement error. The current study focuses on the assessment of language dominance, specifically analyzing one of the most common tools for assessing bilingual language dominance: the Bilingual Language Profile (BLP; Birdsong et al., 2012). While some evidence has been presented for the validity of the interpretation of BLP scores (for construct validity see Gertken et al., 2014; for concurrent validity see Mallonee Gertken, 2013; Solís-Barroso & Stefanich, 2019), it has yet to be examined with respect to reliability.[1] As such, the current study provides an analysis of the test-retest reliability of the BLP, leveraging a large, diverse population of Spanish–English bilinguals, and reporting on the reliability of the language dominance score, the global language scores for the dominant and non-dominant languages, and the individual BLP subcomponents.

2. Literature Review

2.1 Defining Language Dominance

It is worth briefly considering language proficiency as a starting point, as it is both explicit in the subsequent definitions of bilingual language dominance and has been the subject of a long, well-developed body of theoretical and experimental work. Early descriptions considered language proficiency as a speaker's overall competence in a given language (e.g., Thomas, 1994) or as a bipartite concept encompassing both linguistic knowledge (e.g., syntax, lexicon, phonology) and skills (i.e., speaking, listening, reading, writing) (e.g., Carroll, 1972). In light of the relatively decontextualized nature of these definitions of proficiency, others have sought to incorporate the

communicative functions of language (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980). For example, Hulstijn (2011) suggested that language proficiency includes both the core components of "phonetic-phonological, morphonological, morphosyntactic, and lexical domains" (p. 242), and peripheral components such as interactional abilities, strategic competencies, and knowledge of a variety of different types of discourses. While the core components are required for most communicative interaction, the peripheral components are employed selectively depending on the communicative situation (see Hulstijn, 2015). Given the complex and multi-faceted nature of proficiency, as well as the different goals of proficiency measurements in different settings, the operationalization and measurement of proficiency has taken many different forms (for a review see Gaillard & Tremblay, 2016). Among the most common measures of proficiency, as detailed by Olson (2022b), are standardized testing (e.g., Test of English as a Foreign Language [Educational Testing Service, 2020]), self-assessment (e.g., Language Experience and Proficiency Questionnaire [LEAP-Q] [Marian et al., 2007]), single-component tests (e.g., mean length utterance [Baker-Smemoe et al., 2014]), institutional or descriptor-based frameworks (e.g., Common European Framework of Reference [Council of Europe, 2001]), oral proficiency interviews (American Council of Teachers of Foreign Languages, 2012), and institution-specific curricular standards (for discussion see Thomas, 1994).

However, two crucial distinctions should be made between language proficiency and language dominance. First, proficiency generally refers to one's linguistic knowledge and skills in a given language, while language dominance refers to the *relative* abilities between their two languages

(Montrul, 2016). Second, language dominance is considered to have a broader scope than language proficiency, incorporating factors beyond proficiency (Montrul, 2016).

Turing first to the relative nature of language dominance, proficiency generally refers to one's linguistic knowledge and skills in a given language and is most commonly measured in only one language (Montrul, 2016). As such, proficiency is often measured in the L2, particularly in research in second language acquisition, and proficiency in the first language (L1) is assumed to be native-like and stable. In contrast, language dominance refers to a bilingual's relative abilities in each of their two languages. For example, Birdsong (2014) referred to language dominance as the "observed asymmetries of skill in, or use of, one language over the other" (p. 374). Similarly, Treffers-Daller (2019) noted that "language dominance is most often interpreted as referring to the relative strength of a bilingual's proficiency in each language" (p. 379), while Kootstra and Doedens (2016) defined language dominance as a "measure of bilinguals' personal experience with both languages" (p. 711). Many authors have observed that perfectly balanced bilinguals, with equal "mastery" of their two languages, are exceedingly rare (Wei, 2000, p. 6; Romaine, 1999; Treffers-Daller, 2016; among many), if not impossible. More commonly, bilinguals are more dominant in one language and less dominant in the other. While proficiency and dominance usually correlate, this is not necessarily the case. Highlighting the distinction between the relative measure of language dominance and the absolute measure of proficiency, two bilinguals could be "balanced" in their dominance (i.e., roughly equally abilities in both languages), but one could have high proficiency in both languages while another shows lower proficiency in both (Harris et al., 2006; Treffers-Daller, 2011). In practice, while language dominance is a relative measure, calculation of language dominance often relies on comparison

of two absolute measures (i.e., one for each of a bilingual's languages) (for discussion of different methods of comparison see Birdsong, 2016).[2]

Grosjean (2008) wrote that language dominance is reflective of the complementary principle, which holds that a bilingual's two languages develop in response to their different purposes, domains, and relationships. In this vein, languages that are used with fewer interlocutors may be less "fluent" (Grosjean, 2008, p. 24), and linguistic properties that are seldom used (e.g., stylistic varieties) may be underdeveloped. In short, language dominance represents a relative measure of abilities *between* the two languages, while language proficiency is an *absolute* measure in a single language.

The second key difference between proficiency and dominance relates to the scope of the terms. While proficiency is limited to knowledge and skills, dominance is broader, incorporating additional factors. Several authors note two main components of language dominance: language proficiency and language use (for review see Treffers-Daller, 2019). Moreover, language use can be divided into "how frequently bilinguals use their languages" (p. 378) and "how these are divided across domains" (p. 378) such as work, home, and school (Treffers-Daller, 2019). Beyond proficiency and use, others include "individual or environmental factors" (e.g., Martin et al., 2020), biographical factors such age of acquisition and language of education (e.g., Marian et al., 2007; for discussion see Montrul, 2016), context of acquisition (see Martin et al., 2020), and issues of identity and/or attitudes (Birdsong et al., 2012). For example, when evaluating dominance, Birdsong et al.'s (2012) questionnaire assessed proficiency, language use, language history (i.e., linguistic biographical variables), and language attitudes in each of a bilingual's two

languages. In addition, many of these factors have been shown to correlate with bilingual performance (i.e., proficiency). For example, Unsworth (2016) found that experiential variables, like language exposure, correlate with proficiency, although subsequent work suggests that language use might be a stronger predictor (Unsworth et al., 2018). While some have suggested that the definitions of dominance remain underspecified (Cantone et al., 2008; Gertken et al., 2014), Martin et al. (2020) described a degree of conceptual consensus around the notions of language proficiency, language use, and environmental and individual factors, as components of a comprehensive measure of language dominance. Thus, proficiency forms one component of language dominance, but dominance is a more encompassing construct.

2.2. Measuring Language Dominance

Given the conceptual complexity of language dominance, operationalization and measurement has taken a variety of forms. In their review of language dominance assessments, Solís-Barroso and Stefanich (2019) provided a non-exhaustive list of 19 different methods of language dominance assessment used in previous research. Broadly, these different methods can be divided into objective and subjective measures of language dominance. Objective measures rely on tasks that directly measure performance, either written or spoken, individually in each of a bilingual's two languages. The performance is compared between the two languages, with better performance indicating the dominant language. Objective measures noted by Solís-Barroso and Stefanich (2019) included, lexical tasks (e.g., Boston Naming Task [Gollan et al., 2012]), morphosyntactic knowledge tests (e.g., Bedore et al., 2012), semantic knowledge test (e.g., Bedore et al., 2012), oral proficiency (e.g., Gollan et al., 2012), lexical richness (e.g., Treffers-Daller, 2011), and mean length utterance (Yip & Matthews, 2006), among others. Treffers-Daller

(2019) suggested that vocabulary-dependent measures are among the most common objective measures, as they appear to be "more easily quantifiable" (p. 379) than other language proficiency measures. While objective measures provide direct evidence, they are limited in that they largely fail to account for the broader conceptualization of language dominance (Montrul, 2016) and do not include factors beyond proficiency that are commonly considered as part of language dominance (e.g., language history, language attitudes).

Within subjective measures, self-ratings are among the most predominant. Several self-rating tools have been proposed specifically to measure dominance, including the LEAP-Q (Marian et al., 2007), the Bilingual Dominance Scale (Dunn & Fox Tree, 2009), and the BLP (Birdsong, et al., 2012). While these tools differ somewhat in the subcomponents assessed, they all include some measure of proficiency in a bilinguals two languages and seek to provide a relative measure of dominance using the two scores (e.g., a ratio of the strength of Language A to Language B). A number of researchers have suggested that self-ratings are the most common forms of assessing language dominance (e.g., Gertken et al., 2014) for several reasons. First, self-ratings permit the assessment of multiple components of language dominance (e.g., frequency of use, biographical variables, language attitudes) that may not be adequately assessed by objective measures (Gertken et al., 2014). Second, prior research suggests that self-ratings of language abilities are well-correlated with behavioral measures (for review see Gertken et al., 2014). Finally, self-ratings are practical (Treffers-Daller, 2019), providing quick, easy measures with little specialized training required. However, several studies have suggested that self-ratings may differ between groups of different language pairings (Tomoschuk et al., 2019) or potentially between a bilingual's two languages (Delgado et al., 1999).

Considering the selection of an appropriate measure of dominance, Treffers-Daller (2019) provided several key issues to consider. First, the measurement chosen for language dominance should reflect the needs of the study, and given the wide variety of research (and clinical) needs, there is unlikely to be a single optimal measure of language dominance. Second, the chosen measure should be appropriate for each of the languages under study. Treffers-Daller (2019) provided mean length utterance as an example measure that can differ substantially between two languages, making it inappropriate for some language pairings (for discussion see Allen & Dench, 2015). Third, the nature of the cross-language comparison should be made explicit. Solís-Barroso and Stefanich (2019) described that while many measures of dominance provide a categorical output (i.e., which language is dominant), others provide a more gradient scale. Birdsong (2016) rightly argued that the construct of dominance is inherently gradient, not categorical. Moreover, such gradient comparisons may be made via subtraction (Lang A score – Lang B score) or as a ratio (Lang A score/ Lang B score) (for discussion see Birdsong, 2016). Finally, the chosen measure of language dominance (or its corresponding interpretation) should be both valid and reliable.[3] Validity considers whether a given measure adequately represents the underlying target construct or whether the measure is appropriate for a given usage or interpretation, while reliability refers to a measure's demonstrated consistency and repeatability over time. For each of these key issues, the field would benefit from explicit discussion of the appropriateness of a selected measure for a given study, the appropriateness of the measure for a given language pairing or community, and acknowledgement of the degree of reliability and validity of the measure.

As articulation of the proposed or existing uses of an assessment is crucial within an argument-based validity framework (Chapelle, 2011; Kane 2016), it is worth considering the previous usage and interpretation of the language dominance measures. Broadly, language dominance measures have been used to: (1) provide a gradient, relative placement of bilinguals along a dominance continuum (e.g., Amengual & Chamorro, 2015); (2) provide a categorical classification (e.g., dominant in Language A, dominant in Language B, or more balanced) of participants (e.g., Perpiñan, 2018), or (3) provide a screening criteria in which participants who fail to reach a certain dominance threshold are excluded from a given research study (e.g., Gollan et al., 2002). Language dominance scores are often used as a variable of interest (i.e., independent variable) and researchers examine the potential impact of language dominance, either as a relative or categorical (between groups) variable, on a variety of linguistic behaviors. The underlying assumption is that the language dominance measure is representative of an bilingual's underlying language dominance (or relative strength of each language), which has the potential to impact a variety of linguistic outcomes.

2.3. The Bilingual Language Profile (BLP)

2.3.1 Use of the BLP in the Field

The current study focuses on the issue of test-retest reliability for the BLP (Birdsong et al., 2012). The BLP is a self-rated language dominance questionnaire and has been noted as being among the most common subjective measures of language dominance (Solís-Barroso & Stefanich, 2019). As evident from cross-referencing in Google Scholar, the BLP has been cited in hundreds of research papers across a wide range of linguistic (and non-linguistic) subfields, including phonetics and phonology (e.g., Amengual & Chamorro, 2015), morphosyntax (e.g.,

Perpiñán, 2018), lexical acquisition (e.g., Rahman et al., 2018), semantics (e.g., Stocker & Berthele, 2020), speech processing (e.g., Tomé Lourido, 2018), and psycholinguistics (e.g., Poarch et al., 2019), among others.

2.3.2 Design, Scoring, and Interpretation of the BLP

The design of the BLP, detailed fully in Gertken et al. (2014), was conducted in accordance with best practices outlined by Dörnyei and Taguchi (2009) and conceptualizes of language dominance as a multi-faceted construct that places bilinguals along a dominance continuum. The BLP was explicitly designed to respond to potential issues in previous language dominance questionnaires (for discussion see Gertken et al., 2014). Specifically, the BLP was designed to be succinct and easy-to-interpret (c.f., LEAP-Q [Marian et al., 2007]), fully quantitative and intuitive to score (c.f., Bilingual Dominance Scale [Dunn & Fox Tree, 2009]), and easily adaptable to a variety of types of bilinguals in a variety of different communities (c.f., Bilingual Dominance Scale [Dunn & Fox Tree, 2009]; Self-Report Classification Tool [Lim et al., 2008]).

The BLP questionnaire contains 19 questions which are answered for each of a bilingual's two languages or varieties. These questions represent four different subcomponents, each representing a different aspect of language dominance: language history, language use, language proficiency, and language attitudes. Language history (6 questions) collects information about age of acquisition, age at which participants felt comfortable speaking each language, the number of years that participants have spent in a school, country/region, family, and work environment where each language is spoken. The language use (5 questions) subcomponent collects information on the percentage of time, in an average week, that participants currently use

each of their two languages with family, with friends, at work, when talking to themselves, and when counting. The language proficiency (4 questions) subcomponent asks participants to rate their abilities (i.e., "how well do you") in each language across the four language skills – speaking, listening, reading, and writing. Finally, the language attitudes (4 questions) subcomponent asks participants to what degree they feel like themselves when they speak each language, how much they identify with cultures that speak each language, how important it is for them to use each language like a native speaker, and how important it is for them to be perceived as a native speaker of each language. While questions are grouped into four underlying subcomponents, each question is a single-construct item. As such, it is not necessarily the case that responses to each question in a given subcategory will be correlated. Consider, for example, the category of language proficiency. While for many bilinguals, their abilities are closely correlated in each of the four language skills (i.e., reading, writing, speaking, and listening), prior research has shown that heritage speakers often report and perform better in aural receptive competence relative to production and written competencies (Montrul, 2011). As such, while responses within each subcategory may correlate, this is not necessarily the case. Gertken et al., (2014) explicitly acknowledged this issue, noting that the BLP "by taking into account various contexts of language experience, while still providing an overall (context-independent) dominance assessment, is a fair representation of dominance that meets our criteria of efficiency and practicality" (p. 212).

The BLP is quantitatively scored to create a language dominance score (the scoring procedure is detailed in Birdsong et al., 2012). First, a subcategory score is calculated for each subcategory in each language. The subcategory score is determined by summing the raw response for each item

in each subcategory.[4] The subcategory score is then multiplied by a weighting coefficient to provide equal weight to each subcategory score. Table 1 illustrates the weighting coefficient for each category. The global language score is calculated by adding each of the weighted subcategory scores, resulting in a theoretical range between 0 and 218, with 0 corresponding to a complete lack of knowledge and experience with a given language and 218 to a maximal knowledge and experience. Finally, the language dominance score is determined by subtracting the global language score in language A from the global language score in language B, resulting in a continuous dominance score ranging from –218 to 218. The endpoints of the continuum represent maximal dominance in either language A or language B, while 0 represents a 'balanced' bilingual. In interpreting the scores of the BLP, it is worth noting that Birdsong et al. (2012) do no suggest any particular cut-off points (although 0 represents an inflection point from dominance in language A to B), with the continuous nature of the dominance score suggesting that the relative position of two (or more) participants on the scale is of particular relevance.

Table 1. BLP Subcategories and Weighting Coefficients

Subcategory	Scale	Number of Questions	Weighting Coefficient	Total Weight (%)
Language History	0–20	6	0.454	25
Language Use	0–10	5	1.09	25
Language Proficiency	0–6	4	2.27	25
Language Dominance	0–6	4	2.27	25

2.3.3 Previous Validity Studies

To date, a few studies have assessaed the construct and concurrent validity of the BLP (Gertken et al., 2014; Mallonee Gertken, 2013; Solís-Barroso & Stefanich, 2019). Construct validity specifically refers to the appropriateness of interpretations of an underlying theoretical construct resulting from a given measure or the "extent to which a test measures some theoretical

construct" (Byrd & Buckhalt, 1991, pp. 121–122). Considering the construct validity of the BLP, Gertken et al. (2014) reported on an analysis by Amengual and colleagues (Amengual et al., in preparation, as cited in Gertken et al., 2014) who conducted a factor analysis on BLP scores for 68 French–English bilinguals, that indicated "desirable component groups and reflected the underlying dimensions of dominance" (p. 218). Considering the concurrent validity, or degree of agreement between a given measure and previous measures, Solís-Barroso and Stefanich (2019) measured language dominance in 29 Spanish–English bilinguals using five different dominance measures: the BLP (Birdsong et al., 2012), the Bilingual Dominance Scale (Dunn & Fox Tree, 2009), self-ratings of verbal abilities (Flege et al., 2002), self-ratings of written abilities (Flege et al., 2002), and a repetition task (Flege et al., 2002).[5] Results demonstrated a moderate correlation between the BLP and the Bilingual Dominance Scale, and a strong correlation between the BLP and self-ratings of verbal abilities. No significant correlations were found between the BLP and either self-ratings of written abilities or the repetition task. Similarly, Mallonee Gertken (2013) reported on the correlation between the BLP and two objective proficiency measures for 65 French–English bilinguals: the Oxford Placement Test (OPT) and a cognitive naming task. The OPT is a multiple-choice test focusing on lexical, grammatical, and pragmatic knowledge. Results showed a strong correlation between the self-rated BLP French proficiency scores and performance on the French OPT. Results, for a subset of participants, also showed a moderate correlation between the cognitive naming task in French and the BLP dominance score. Taken collectively, these studies provide initial evidence for the validity of the BLP, suggesting that it provides an appropriate measure of the underlying construct of language dominance and correlates well with performance on other subjective and objective measures.

2.4 Research Questions

Given the important role that language dominance plays in the field of bilingualism, carefully designed methods of assessment are crucial for advancing theory. Recent work has directly called for improved methodologies for examining bilingual language experiences (de Bruin, 2019) and highlighted the wide variability (Treffers-Daller, 2019) and potential pitfalls of current measures of language dominance (Solís-Barroso & Stefanich, 2019). To assess the usefulness of a given measure of language dominance, two key components are necessary: validity and reliability. Validity broadly refers to "the extent to which a psychometric instrument measures what it has been designed to measure" (Dörnyei & Taguchi, 2009, p. 93) or whether an instrument provides an appropriate evaluation (Kane, 2016). Reliability refers to "the extent to which scores on the instrument are free from errors of measurement" (Dörnyei & Taguchi, 2009, p. 93). Test-retest reliability, the approach taken in the current study, specifically addresses the stability of a given measure over time, with greater consistency in the measure over time corresponding to less measurement error. In the case of the BLP, evidence has been presented for both internal (Gertken et al., 2014) and external (Mallonee Gertken, 2013) validity, and the BLP may be considered to be appropriate for the previously specified uses (e.g., providing a relative measure of participants' language dominance along a continuum). Yet, while the BLP has gained significant traction among researchers in the field, it has yet to be systematically assessed with respect to reliability. Addressing this key methodological gap in the field, the current paper has three specific research aims.

The first aim is to assess the test-retest reliability of the BLP's measure of language dominance. Given previous research that has suggested differences in self-rating abilities between a

bilingual's two languages (Delgado et al., 1999), the second aim is to assess the test-retest reliability of the global language score in both the dominant and non-dominant languages. As a corollary to this second research aim, this study examines whether the global language score is more reliable in one language or the other. Finally, this study examines the test-retest reliability of each of the individual subcomponents of the BLP (i.e., language history, language use, language proficiency, language attitudes) and compares the reliability of the subcomponents.

To address the above research aims, a test-retest approach to reliability was employed. Test-retest methods provide an estimate of the reliability or stability of a particular measure or construct over time. Broadly, the more comparable scores are between the initial and follow-up testing sessions, the more reliable the measure. To conduct the test-retest reliability analysis, the BLP was administered to a large, varied sample of Spanish–English bilinguals and a second follow-up survey was administered approximately one month later. Analyses focus on the reliability of the overall language dominance score, the dominant and non-dominant global language scores, and each of the subcomponent scores.

3. Methods

3.1 Participants

Initial recruitment targeted Spanish–English bilinguals, colloquially defined in recruitment materials as "anyone who can comfortably carry out daily conversations in English and Spanish." The decision to recruit bilinguals with a moderate (or greater) degree of fluency or proficiency paralleled previous research on the BLP (Solís-Barroso & Stefanich, 2019). Moreover, in line with a broad definition of bilingualism (e.g., Montrul, 2016), recruitment

materials specifically noted that "it does not matter at what age you learned these languages" or "in which language you feel most comfortable." To provide a well-rounded sample, participants were recruited from a wide range of ages, ethnic backgrounds, origins, and geographic regions. Participants were recruited online via snowball sampling (n =303) and through the crowd-sourcing platform Prolific (n =151).[6] Online crowd-sourcing platforms, such as Prolific, have been shown to be a reliable method of collecting high-quality data (Hauser & Schwarz, 2016). This reliability has been extended to collection of Spanish-language data for research in linguistics (Nagle, 2019; Ortega-Santos, 2019). As the initial snowball sample skewed towards English-dominant speakers, additional participant inclusionary criteria (i.e., native Spanish speaker, also speaks English) were used to recruit participants in Prolific. In addition, following recommendations by Peer et al. (2014), inclusionary criteria of prior approval rate and number of previous submissions were used to ensure the quality of responses. All participants received compensation for their participation in the study.

Of the initial 454 participants, 422 consented to be contacted for a follow-up survey (approximately 93%). Of those, 283 completed the second survey (67%). An additional question was included in the survey to establish any potential changes in a participant's daily life that could result in changes in their patterns of language use (e.g., moving to a new city).[7] Thirty-five participants reported a life change that could impact patterns of language use and were eliminated from the analysis.

A total of 248 (female = 156, male = 89, non-binary, trans, or no response = 3), ranging in age from 18–75 (M = 29.8, SD = 10.7), were retained for the final analysis. Considering ethnic

background, participants were able to select more than one background, resulting in a total of 287 ethnic background tags. A majority of participants identified as Hispanic, Latino, or Spanish origin (Table 2). Considering origin for participants identifying as Hispanic, Latino, or Spanish, again, participants were able to provide more than one origin (e.g., Mexican–Guatemalan). As illustrated in Table 3, the majority of participants identifying as Hispanic, Latino, or Spanish provided Mexican as their origin. The predominance of Mexicans parallels the Hispanic population in the United States, where the majority of participants were based (Noe-Bustamante et al., 2019). Finally, considering geographic location, a majority (n = 243) of participants were from the United States, with states that have large Hispanic populations (e.g., California, Texas, Florida) well-represented in the data (U.S. Census, 2020). The geographic distribution of participants is illustrated in Figure 1.

Table 2. Participant Ethnic Background.

Ethnic Background	***n***
Hispanic, Latino, or Spanish Origin	208
White	64
Asian	5
American Indian or Native Alaskan	4
No response	4
Black or African American	2

Table 3. Origin for Participants Identifying as Hispanic, Latino, or Spanish (n > 5).

Hispanic, Latino, or Spanish Origin (please specify)	***n***
Mexican	94
Spanish	13
Colombian	11
Puerto Rican	11
Venezuelan	11
Peruvian	10
Dominican	9
Costa Rican	7
Argentine	6
Salvadoran	6
Cuban	6
Ecuadorian	6

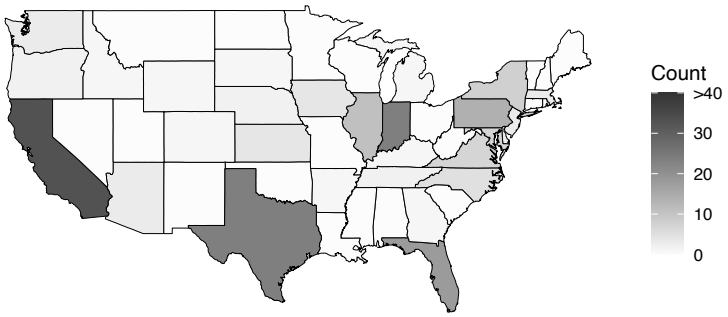

Figure 1. Geographic distribution of study participants in the United States.

3.2 Procedure

Participants were able to select the language in which they preferred to complete the questionnaire (English n = 187; Spanish n = 61). Each participant completed two online questionnaires during a single session: the Bilingual Language Profile (Birdsong et al., 2012) and the Bilingual Code-switching Profile (Olson, 2022a), along with several open-ended questions. The current study focuses only on the responses to the BLP. The median time to complete all surveys was approximately 17.4 minutes.[8] The estimated time to complete the BLP was 9–11 minutes.

Embedded within the two questionnaires, four different quality checks were included to ensure that participants sufficiently engaged with the material. Two attention-check questions requested a specific response from participants (e.g., "how many years have you… please mark the number seven for your answer"). Two language-oriented checks consisted of the same factual biographical questions presented in English (e.g., age) and Spanish (e.g., *edad*) at different points in the survey. Responses were compared between the related questions, and identical responses

were required to pass the language-oriented quality checks. No participant failed two or more quality checks (see Berinsky et al., 2013), and all were retained for the subsequent analysis.

Participants who consented to a follow-up survey were contacted by email a minimum of four weeks after the completion of the initial questionnaire. The four-week time interval was selected as it minimized potential memory effects, while limiting the degree of expected change in language dominance (for discussion of test-retest intervals see Chmielewiski & Watson, 2009). The mean interval between the completion of the first and second questionnaires was approximately 1 month (M = 32.4 days, SD = 8.8 days). The second questionnaire was identical to the first, with the exception of different quality checks.

3.3 Analysis

Statistical analysis was conducted using R (R Core Team, 2021). Test-retest reliability was evaluated via an intraclass correlation (ICC), using the *irr* package (Gamer et al., 2019), with a single-measurement, absolute agreement, two way mixed-effects model (for selection of different ICC forms see Koo & Li, 2016). Comparisons of ICC values were conducted by generating a bootstrapped distribution of ICC values, using the *boot* package (Canty & Ripley, 2021), and analyzing mean differences and confidence intervals.

4. Results

4.1 Reliability of Overall BLP Dominance Score

Given that, theoretically the BLP dominance score can range from –218 (Spanish-dominant) to +218 (English dominant), an examination of the overall dominance scores at Time 1 (T1)

suggests a wide range of dominance profiles (range = –116.1 to 183.0), with a slight skew towards English-dominance (*M* = 17.9, *SD* = 61.7). Figure 2 illustrates the distribution of the participants across the language dominance continuum. The overall skew of the data towards English dominance is not surprising, given that a large majority of participants currently resided in the United States, where English functions as the majority language in most regions and communities. An analysis of the data as a whole suggests a wide-ranging and varied sample of bilingual dominance profiles.

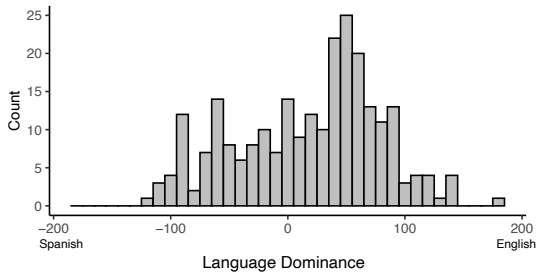

Figure 2. Histogram of language dominance scores (T1).

To initially examine the relationship between the BLP dominance scores produced by participants at the test (T1) and retest (Time 2 [T2]) sessions, an intraclass correlation was conducted, with a single measurement, absolute agreement, two way mixed-effects model. Results of the intraclass correlation demonstrated an "excellent" level of test-retest reliability (ICC(A,1) = .979, 95% CI [.973, .984]) (Koo & Li, 2016).[9] The comparison of dominance scores at T1 and T2 is illustrated in Figure 3. Highlighting the test-retest reliability, Figure 4 illustrates the Bland-Altman plot (Bland & Altman, 1986), depicting a participant's average score over T1 and T2 relative to the difference in their scores between T1 and T2. The overall mean difference between scores at T1 and T2 was *M* = 1.84 (*SD* = 12.37). The grand average of mean scores between T1 and T2 was *M* = 16.94 (*SD* = 60.53). Importantly, difference scores were uniformly

distributed across the full range of average dominance scores, suggesting that the BLP is equally reliable for bilinguals from a wide range of dominance profiles.

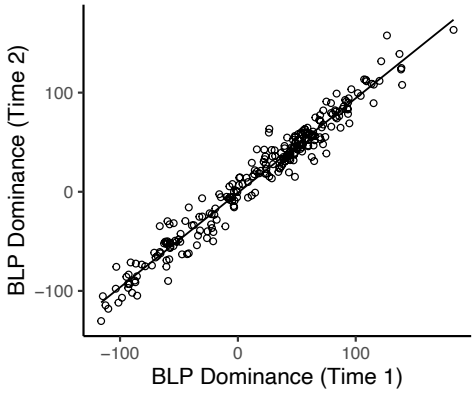

Figure 3. Scatter plot of BLP dominance score at T1 and T2.

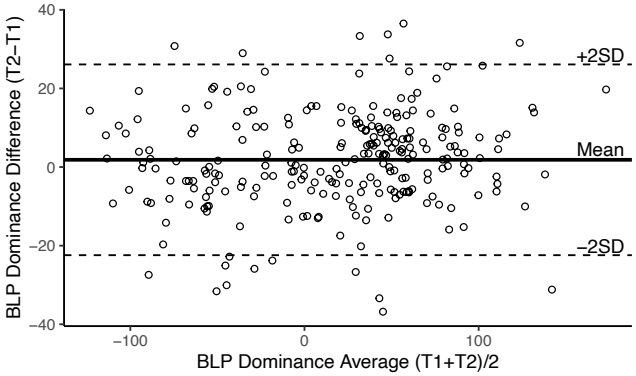

Figure 4. Bland-Altman plot of BLP dominance scores.

4.2 Reliability of Global Language Scores

As the language dominance score was computed by first calculating a global language score in both English and Spanish and then calculating the difference, and given previous research that has suggested that bilinguals may be more accurate in reporting behavior in one of their two

languages (Delgado et al., 1999), it was relevant to examine the test-retest reliability of the global language scores in a participant's dominant and non-dominant languages. First, a participant's language dominance was determined by examining the dominance score at T1. Dominance scores greater than 0 indicated English as the dominant language. Dominance score less than 0 indicated Spanish as the dominant language. The opposite language was the non-dominant.[10]

Considering the test-retest reliability of the global language score in the dominant language, an intra-class correlation was conducted. Results of the intraclass correlation (single measure, absolute agreement, two-way mixed effects model) showed good reliability (ICC(A,1) = .890, 95% CI [.859, .914]) of the dominant global language score. Considering the test-retest reliability of the global language score in the non-dominant language, parallel analysis suggested overall excellent reliability (ICC(A,1) = 0.919, 95% CI [.898, .937]). Figure 5 illustrates the relationships between the global language scores at T1 and T2 in each language. Differences in the distributions of the overall scores, with the dominant language scores generally distributed at the higher end of the scale relative to the non-dominant language scores, are generally to be expected. Figure 6 depicts a pair of Bland-Altman plots for the dominant and non-dominant languages, comparing the global language score difference with the global language score average by participant. Again, analysis of Figure 6 shows that difference scores were distributed consistently across the full range of average dominance scores, highlighting that in both the dominant and non-dominant languages, the BLP is reliable across a range of global language scores. Taken as a whole, the statistical and visual analyses show that the BLP global language

score demonstrates strong test-retest reliability in both the dominant and non-dominant languages.

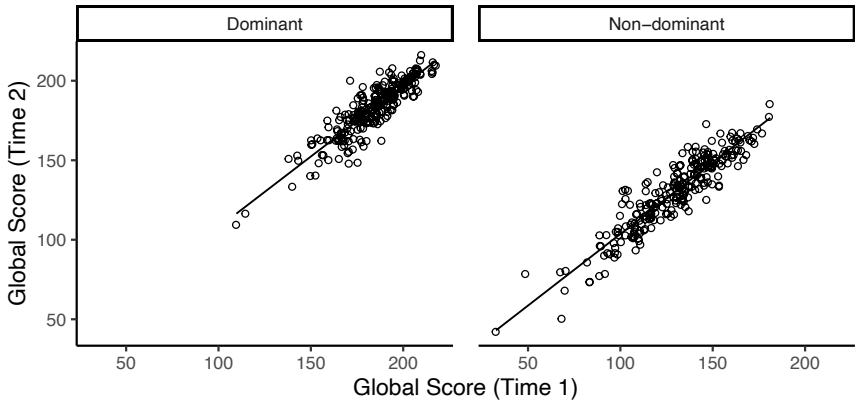

Figure 5. Scatter plot of the global language scores for dominant (left) and non-dominant languages (right) at T1 and T2.

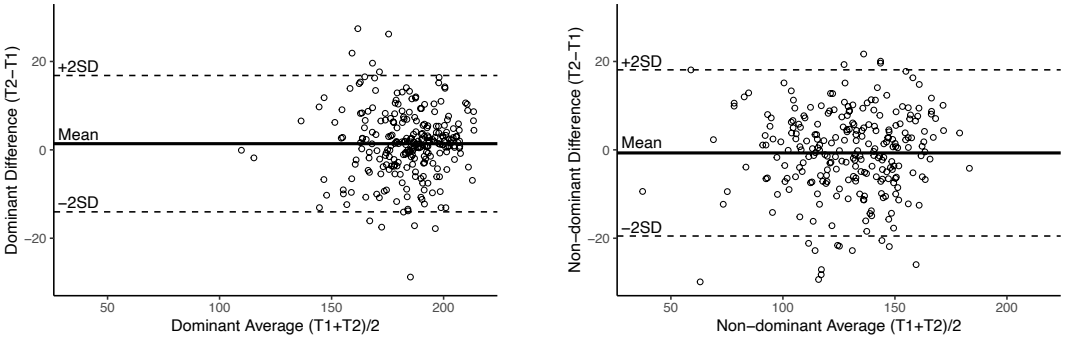

Figure 6. Bland-Altman plots of global language scores for the dominant (left) and non-dominant languages (right).

Finally, the test-retest reliability of the dominant and non-dominant global language scores were compared using a bootstrap resampling method. Specifically, a bootstrapped distribution of ICC values was generated for both dominant (M_{ICC} = 0.888, SD = 0.017) and non-dominant global language scores (M_{ICC} = 0.918, SD = 0.011), using the *boot* package (Canty & Ripley, 2021) with 1000 iterations. A 95% CI for the difference in the means was then calculated (M_{diff} = -

0.0297, 95% CI [-0.0310, -0.0285]). Results show a significant difference (i.e., 95% CI does not contain 0) between the bootstrapped ICC values for the dominant and non-dominant global language scores, with the non-dominant scores showing greater reliability than the dominant global language scores.

4.3 Reliability of Subcomponent Scores

To examine test-retest reliability of each of the subcomponents of the BLP score, a series of intraclass correlations were conducted. To provide an overall understanding of the reliability by subcomponent, data was pooled for responses in the dominant and non-dominant languages. Results are available in Table 4, and Figure 7 illustrates the relationship between each weighted subcomponent score at T1 and T2. Taken as a whole, each individual subcomponent demonstrated "good" to "excellent" reliability (Koo & Li, 2016, p. 158).

Table 4. Reliability by Subcomponent.

Subcategory	ICC(A,1)	95% CI
Language History	0.926	[.913, .938]
Language Use	0.960	[.952, .966]
Language Proficiency	0.881	[.860, .900]
Language Attitudes	0.858	[.833, .880]

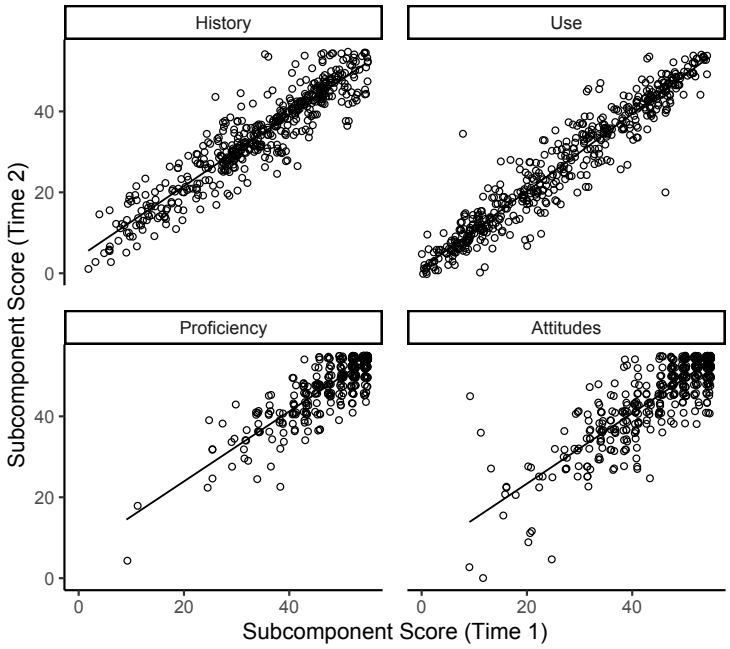

Figure 7. Scatter plot of BLP subcomponent scores at T1 and T2.

In the analysis of ICC values, some differences can be observed among the subcomponents, with language use as showing the strongest test-retest reliability, and language proficiency and language attitudes with somewhat lower reliability. To assess whether these differences were statistically significant, a bootstrap method was again employed to create distributions for the subcomponents' ICC values (1000 iterations). A series of pairwise comparisons were conducted by calculating the mean differences and Bonferroni-adjusted 95% CIs for each pair of subcomponents. Results (Table 5) showed significant differences between each pair of subcomponents. ICC values for language use were found to be the highest, followed by language history, language proficiency, and language attitudes.

Table 5. Comparison of Subcomponent ICC Values.

Comparison	Boot Mean Diff	95% CI
Language History – Language Use	-0.0339	[-0.0345, -0.0334]
Language History – Language Proficiency	0.0453	[0.0437, 0.0468]
Language History – Language Attitudes	0.0688	[0.0670, 0.0706]

Language Use – Language Proficiency	0.0792	[0.0777, 0.0806]
Language Use – Language Attitudes	0.1027	[0.1010, 0.1044]
Language Proficiency – Language Attitudes	0.0235	[0.0214, 0.0257]

5. Discussion

The main goal of the current study was to examine the test-retest reliability of the BLP and its various subcomponents. The overarching findings suggest excellent test-retest reliability of the BLP language dominance score, and highlight its appropriateness for providing a gradient and relative measure of language dominance. In examining the reliability of the global language scores and various subcomponents, all were found to evidence good to excellent test-retest reliability for this particular population. However, some small, but significant, differences emerged between the reliabilities of the dominant and non-dominant global language scores, as well as the reliability of the different subcomponents. The discussion focuses on these differences.

5.1 Comparing the Reliability of Dominant and Non-dominant Global Language Scores

First, the BLP global language scores, for both the dominant and non-dominant languages, demonstrated good to excellent degrees of test-retest reliability. These findings provide further evidence for the use of the BLP as reliable measure of language dominance. Yet, it may be of interest to consider why reliability was greater in the non-dominant language than the dominant language, although it should be noted that the magnitude of the difference in reliability was small (M_{diff} = -0.0297). Two possible explanations are considered here, both theoretical and methodological.

From a theoretical perspective, a participant's ability to reliably (i.e., consistently) respond to questions may be directly related to the degree of conscious awareness that they have of their own language abilities. Research in language awareness, defined as the "explicit knowledge about language, and conscious perception and sensitivity in language learning… and language use" (Association for Language Awareness, n.d.), has shown that several techniques enhance language awareness. These techniques, including analytical discussions about language and verbalizing ideas about language (van der Broek et al., 2022), have been shown to impact both cognitive (e.g., awareness of language structures and communicative functions) and affective (e.g., forming language attitudes) levels (Farias, 2005). Language awareness building techniques are present in many second language classrooms and pedagogical materials, potentially leading to greater awareness of one's abilities in the non-dominant language. Considering previous research, Delgado et al. (1999) examined the correlation between bilinguals' proficiency self-ratings in English and Spanish and their performance on an objective measure of proficiency via the Woodcock-Muñoz test (Woodcock & Muñoz-Sandoval, 1993). They found that participants were more accurate in self-rating Spanish abilities relative to English abilities. They speculated that participants, all currently living in the United States, likely had taken "foreign" language classes in Spanish, in which they received direct feedback regarding their Spanish skills, effectively raising their awareness of their Spanish-language skills. Applied to the current study, while the data is not available on language classes, many participants may have taken conscious steps towards improving skills in their non-dominant language, such as language courses, language learning apps, or various forms of self-study. As such, much like participants in Delgado et al. (1999), participants in the current study may have received feedback about, and have greater awareness of, their non-dominant language abilities. This greater awareness may

translate directly into a greater reliability in test-retest measures for the non-dominant language relative to the dominant language.

From a methodological perspective, it should be noted that global language scores in the dominant language cluster towards the top end of the range, while scores in the non-dominant language evidence greater dispersion (see Figure 5). As noted by Lehmann (2007, p. XX), native competence (i.e., dominant language) is "typically closer to perfection (thus to a pole of the assessment scale)" thus suffering from a degree of ceiling effects. In contrast, measures in the non-dominant language may evidence greater spread across the full range of global score values. In this case, the natural restriction of ranges in dominant language abilities, and as a result in the dominant global language score, may serve to attenuate correlations (Fife et al., 2012) relative to the non-dominant language. Given that this range restriction is inherent in a bilingual's dominant language, this effect may impact most relative measures of language dominance (and proficiency).

5.2 Comparing the Reliability of the BLP Subcomponents

Again, in any discussion of the results by subcomponent, it should be noted that the overall results for each subcomponent illustrate a high degree of test-retest reliability. Differences in the subcomponents' reliability scores, while interesting from a theoretical perspective, should not be taken as an inherently negative evaluation of the reliability of the measure as a whole. That said, it is worth considering several possible explanations that may impact, either individually or collectively, the relative reliabilities of each of the four BLP subcomponents (language history, language use, language proficiency, and language attitudes). An analysis of the results in Section

4.3 highlight differences between each of the subcomponent reliability scores, with language use and language history evidencing the highest test-retest reliabilities and language attitudes and language proficiency the lowest reliabilities. These differences may be attributed to the overall difficulties in measuring the constructs represented by each subcomponent or the general malleability of each construct.

Considering the overall difficulty in measuring each construct, there are clear differences between the more and less reliable subcomponents. Specifically, language use and language history measure concrete, easy-to-conceptualize components. The link between question and construct is transparent. In contrast, language attitudes and language proficiency are psychological constructs, and the link between specific questions and the underlying construct is more opaque. For example, in discussing language attitudes, Garrett (2010) noted that "the status of attitudes as psychological constructs brings difficulty in accessing them" (p. 20), resulting in debate about the effectiveness of different attitude measures. Moreover, direct measures of attitudes, such as those employed here, may be susceptible to bias (see Schleef, 2022). In short, differences in the reliability scores between subcomponents may be driven, in part, by the nature of each underlying construct and the relative difficulty (or ease) of operationalizing and measuring those constructs. A second possible distinction between the subcomponents is the inherent variability of each subcomponent. With respect to language history, given that questions refer to factual events in a participant's life (e.g., how many years have you…), responses to such questions are highly unlikely to change between the first and second iterations of the questionnaire. Similarly, as participants who experienced any major life changes that could impact their daily language use were removed from analyses, responses to language use

questions were likely to be inherently stable. In contrast, several authors have noted that, language attitudes are inherently variable, responding to social, psychological, and political pressures (Satraki, 2019). In the current data, this appears to be particularly relevant for participants who have overall low to mid attitude ratings (see Figure 7). Some have noted that language attitudes may change "moment to moment", although such systematic variation is not "entirely contradictory to the idea of durability" (Garrett, 2010, p. 30). Thus, some subcomponents (e.g., language attitudes) may be inherently more variable than others (e.g., language history), and evidence greater degrees of change between the first and second tests, resulting in differences in the subcomponent reliabilities.

6. Conclusion

Language dominance has been a crucial factor in examining bilingualism in both research and clinical settings (e.g., Gertken et al., 2014). A wide range of methods have been proposed for measuring language dominance (for review see Solís-Barroso & Stefanich, 2019), and the Bilingual Language Profile (Birdsong et al., 2012) has been one of the most frequently used assessment tools. While previous studies have begun to assess the validity of the BLP (Gertken et al., 2014; Mallonee Gertken, 2013; Solís-Barroso & Stefanich, 2019), it had yet to be systematically evaluated with respect to reliability. The present study had as its overarching aim to assess the test-retest reliability of the BLP as a measure of a bilingual's language dominance. As a corollary, this study also examined the relative reliability of the BLP in a bilingual's dominant and non-dominant languages, and among the BLP's subcomponents. Findings from the current study, coupled with previous research showing that the BLP demonstrates a degree of construct validity (Gertken et al., 2014) and correlates well with both other objective and

subjective measures of language dominance (Mallonee Gertken, 2013; Solís-Barroso & Stefanich, 2019), suggest that the BLP may be valid and reliable method of assessing language dominance.

Although the current study suggests strong test-retest reliability for the BLP, it should be acknowledged that these findings are limited in several specific ways. As recruitment was limited to bilinguals who were fairly proficient in both languages, parallel to previous research on the BLP (Mallonee Gertken, 2013; Solís-Barroso & Stefanich, 2019), future research should examine bilinguals from the full proficiency and dominance spectra, including L2 learners. Similarly, this research was limited to Spanish–English bilinguals mainly residing in the United States. A more holistic examination of the reliability and validity of the BLP would benefit from incorporating different language pairings (for influence of language on self-rating see Tomoschuk et al., 2019) that have substantially different social statuses, such as diglossic contexts. While the construct of language dominance is multifaceted and potentially difficult to operationalize (e.g., Marian et al., 2019), researchers should aim to choose dominance assessment methods that have been assessed for validity and reliability. Although Treffers-Daller (2019) wrote that there is unlikely to be a single optimal measure, given the variety of research aims and needs, evidence has begun to suggest that the use and interpretation of scores from the BLP may be valid for their intended lower-stake and diagnostic uses (Gertken et al., 2014; Mallonee Gertken, 2013; Solís-Barroso & Stefanich, 2019) and a reliable (current study) measure of assessing bilingual dominance, and appropriate for describing participants' relative dominance along a continuum. Moving forward, researchers should continue to evaluate the methods used for assessing language dominance and work to promote tools that have been

shown to be reliable and valid in their use and interpretation, effectively serving to enhance the comparability of research from across the field.

Notes

[1] As noted by an anonymous reviewer, within an argument-based approach to validity, one may consider strong reliability as a necessary condition for valid use and interpretation of a given test or measure. In that case, reliability may be interpreted as a component of validity. Acknowledging this perspective, reliability, either as a component of validity or as a property of a measure, remains an important element in the evaluation of any language assessment.

[2] In the BLP (Birdsong et al., 2012), the global language scores provide an absolute measure which are then used to compute the language dominance score.

[3] While Treffers-Daller (2019) specifically mentioned validity, both validity and reliability were key components in the psychometric assessment.

[4] Questions regarding the age at which participants began learning a given language and began to feel comfortable using a given language were reverse scored. Responses of "since birth" and "as early as I can remember" are worth 20 points. Responses of "20 +" and "not yet" received 0 points.

[5] In the repetition task, participants orally repeated utterances in each language that were presented auditorily. Dominance in the repetition task was calculated based on ratio of sentence production duration (in milliseconds) in each language, with an "underlying assumption" (p. 9) that speakers produce sentences faster in the dominant language relative to the non-dominant language (Solís-Barroso & Stefanich, 2019).

[6] Web-based crowd-sourcing platforms, such as Prolific, connect researchers to remotely-based participants who complete a given task, in this case the questionnaires, for compensation. Nagle (2019) explained these platforms allow access to a diverse population without some of the practical barriers associated with in-person recruitment.

[7] The question regarding life changes was: "Since your response to our first survey, have you experienced any important changes in your life that could impact your language use? (For example: moving to a new city; changing jobs; changing family structure through a new relationship, separation, death; etc.)." Response options were: Yes/No. While any participant selecting "yes" was eliminated from analysis, an open-ended follow-up question was also included: "If you answered 'yes', please explain."

[8] The mean time to completion was 22.7 minutes (*SD* = 23.2). Median time to completion likely represents a better estimate, as completion time data is positively skewed, with a maximum time to completion of 3.5 hours. Given the online, self-paced procedure, it is likely that some participants did not complete the survey without interruption.

[9] In contextualizing ICC estimates, Koo and Li (2016) suggested that values less than .5 may be interpreted as indicating "poor" reliability, between .5 and .75 as "moderate", between .75 and .9 as "good", and above .9 as "excellent" (p. 158).

[10] While the dominant and non-dominant languages were determined using the dominance score at T1, there was wide agreement between dominance at T1 and T2. Of the 248 participants, overall language dominance differed between T1 and T2 for only 2.4% of the participants (*n* = 6).

References

Allen, S. E., & Dench, C. (2015). Calculating mean length of utterance for eastern Canadian Inuktitut. *First Language*, *35*(4-5), 377–406. https://doi.org/10.1177/0142723715596648

Amengual, M., & Chamorro, P. (2015). The effects of language dominance in the perception and production of the Galician mid vowel contrasts. *Phonetica*, *72*(4), 207–236. https://doi.org/10.1159/000439406

American Council of Teachers of Foreign Languages. (2012). *ACTFL Proficiency Guidelines 2012.* https://www.actfl.org/resources/actfl-proficiency-guidelines-2012

Association for Language Awareness. (n.d.). *About.* https://www.languageawareness.org/?page_id=48

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Baker-Smemoe, W., Dewey, D. P., Bown, J., & Martinsen, R. A. (2014). Does measuring L2 utterance fluency equal measuring overall L2 proficiency? Evidence from five languages. *Foreign Language Annals*, *47*(4), 707–728. https://doi.org/10.1111/flan.12110

Bedore, L. M., Peña, E. D., Summers, C. L., Boerger, K. M., Resendiz, M. D., Greene, K., Bohman, T. M., & Gillam, R. B. (2012). The measure matters: Language dominance profiles across measures in Spanish–English bilingual children. *Bilingualism: Language and Cognition*, *15*(3), 616–629. https://doi.org/10.1017/S1366728912000090

Berinsky, A. J., Margolis, M. F., & Sances, M. W. (2013). Separating the shirkers from the workers? Making sure respondents pay attention on self-administered surveys. *American Journal of Political Science, 58*(3), 739-753. https://doi.org/ 10.1111/ajps.12081

Birdsong, D. (2014). Dominance and age in bilingualism. *Applied Linguistics 35*, 374–392. https://doi.org/10.1093/applin/amu031

Birdsong, D. (2016). Dominance in bilingualism: Foundations of measurement, with insights from the study of handedness. In C. Silva-Corvalán & J. Treffers-Daller (Eds.), *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 85–105). Cambridge University Press. https://doi.org/10.1017/CBO9781107375345.005

Birdsong, D., Gertken, L. M., & Amengual, M. (2012). *Bilingual Language Profile: An easy-to-use instrument to assess bilingualism.* COERLL, University of Texas at Austin. https://sites.la.utexas.edu/bilingual/

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, *327*(8476), 307–310. https://doi.org/10.1016/S0140-6736(86)90837-8

Byrd, P. D., & Buckhalt, J. A. (1991). A multitrait-multimethod construct validity study of the Differential Ability Scales. *Journal of Psychoeducational Assessment*, *9*(2), 121–129. https://doi.org/10.1177/073428299100900202

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Cantone, K., Kupisch, T., Müller, N., & Schmitz, K. (2008). Rethinking language dominance in bilingual children. *Linguistische Berichte*, *2008*(215), 307–343.

Canty, A., & Ripley, B. D. (2021). *boot: Bootstrap R (S-Plus) functions*. R package version 1.3-28. https://cran.r-project.org/web/packages/boot

Carroll, J. (1972). Fundamental considerations in testing for English language proficiency of foreign students. In H. B. Allen & R. N Campbell (Eds.), *Teaching English as a second language* (pp. 313–320). McGraw-Hill.

Chapelle, C. A. (2011). Validity argument for language assessment: The framework is simple… *Language Testing, 29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Chmielewiski, M., & Watson, D. (2009). What is being assessed and why it matters: The impact of error on trait research. *Journal of Personality and Social Psychology, 97*(1), 186–202. https://doi.org/10.1037/a0015618

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press.

de Bruin, A. (2019). Not all bilinguals are the same: A call for more detailed assessments and descriptions of bilingual experiences. *Behavioral Sciences*, *9*(33), 1–13. https://doi.org/10.3390/bs9030033

Delgado, P., Guerrero, G., Goggin, J. P., & Ellis, B. B. (1999). Self-assessment of linguistic skills by bilingual Hispanics. *Hispanic Journal of Behavioral Sciences*, *21*(1), 31–46. https://doi.org/10.1177/0739986399211

Dörnyei, Z., & Taguchi, T. (2009). *Questionnaires in second language research: Construction, administration, and processing*. Routledge.

Dunn, A. L., & Fox Tree, J. E. (2009). A quick, gradient bilingual dominance scale. *Bilingualism: Language and Cognition*, *12*(3), 273-289. https://doi.org/10.1017/S1366728909990113

Educational Testing Service. (2020). *Test of English as a Foreign Language.* www.ets.org/toefl

Farias, M. (2005). Critical language awareness in foreign language learning. *Literatura y Lingüística, 16,* 211–222. https://doi.org/10.4067/S0716-58112005000100012

Fife, D. A., Mendoza, J. L., & Terry, R. (2012). The assessment of reliability under range restriction: A comparison of α, ω, and test–retest reliability for dichotomous data. *Educational and Psychological Measurement*, *72*(5), 862–888. https://doi.org/10.1177/0013164411430225

Flege, J. E., MacKay, I. R., & Piske, T. (2002). Assessing bilingual dominance. *Applied Psycholinguistics*, *23*(4), 567–598. https://doi.org/10.1017/S0142716402004046

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, *66*(2), 419–447. https://doi.org/10.1111/lang.12157

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Garrett, P. (2010). *Attitudes to language*. Cambridge University Press. https://doi.org/10.1017/CBO9780511844713

Gertken, L. M., Amengual, M., & Birdsong, D. (2014). Assessing language dominance with the bilingual language profile. In P. Leclercq, A. Edmonds, H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 208–225). Multilingual Matters.

Gollan, T. H., Montoya, R. I., & Werner, G. A. (2002). Semantic and letter fluency in Spanish-English bilinguals. *Neuropsychology*, *16*(4), 562–576. https://doi.org/10.1037/0894-4105.16.4.562

Gollan, T. H., Weissberger, G. H., Runnqvist, E., Montoya, R. I., & Cera, C. M. (2012). Self-ratings of spoken language dominance: A Multilingual Naming Test (MINT) and preliminary

norms for young and aging Spanish–English bilinguals. *Bilingualism: Language and Cognition*, *15*(3), 594-615. https://doi.org/10.1017/S1366728911000332

Grosjean, F. (2008). *Studying bilinguals*. Oxford University Press.

Harris, C. L., Gleason, J. B., & Aycicegi, A. (2006). When is a first language more emotional? Psychophysiological evidence from bilingual speakers. In A. Pavlenko (Ed.), *Bilingual minds: Emotional experience, expression, and representation* (pp. 257–283). Multilingual Matters.

Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better online attention checks than do subject pool participants. *Behavioral Research, 48,* 400–407. https://doi.org/10.3758/s13428-015-0578-z

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, *8*(3), 229–249. https://doi.org/10.1080/15434303.2011.565844

Hulstijn, J. H. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins.

Kane, M. T. (2016). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kootstra, G. J., & Doedens, W. J. (2016). How multiple sources of experience influence bilingual syntactic choice: Immediate and cumulative cross-language effects of structural

priming, verb bias, and language dominance. *Bilingualism: Language and Cognition*, *19*(4), 710–732. https://doi.org/10.1017/S1366728916000420

Lehmann, C. (2007). Linguistic competence: Theory and empiry. *Folia Linguistica, 41*(3–4), 223–287. https://doi.org/10.1515/flin.41.3-4.223

Lim, V. P., Liow, S. J. R., Lincoln, M., Chan, Y. H., & Onslow, M. (2008). Determining language dominance in English–Mandarin bilinguals: Development of a self-report classification tool for clinical use. *Applied Psycholinguistics*, *29*(3), 389–412. https://doi.org/10.1017/S0142716408080181

Mallonee Gertken, S. E. (2013). *Priming of relative clause attachment during comprehension in French as a first and second language.* [Doctoral dissertation, University of Texas at Austin]. https://repositories.lib.utexas.edu/handle/2152/21778

Marian, V., Blumenfeld, H. K., & Kaushanskaya, M. (2007). The Language Experience and Proficiency Questionnaire (LEAP-Q): Assessing language profiles in bilinguals and multilinguals. *Journal of Speech, Language, and Hearing Research*, *50,* 940–967. https://10.1044/1092-4388(2007/067)

Martin, J. M., Altarriba, J., & Kazanas, S. A. (2020). Is it possible to predict which bilingual speakers have switched language dominance? A discriminant analysis. *Journal of Multilingual and Multicultural Development*, *41*(3), 206–218. https://doi.org/10.1080/01434632.2019.1603236

Montrul, S. (2011). Introduction: The linguistic competence of heritage speakers. *Studies in Second Language Acquisition*, *33*(2), 155–161. https://doi.org/10.1017/S0272263110000719

Montrul, S. (2016). Dominance and proficiency in early and late bilingualism. In C. Silva-Corvalán and J. Treffers-Daller (Eds.), *Language dominance in bilinguals: Issues of*

measurement and operationalization (pp. 15–35). Wiley-Blackwell. https://doi.org/10.1017/CBO9781107375345

Nagle, C. (2019). Developing and validating a methodology for crowdsourcing L2 speech ratings in Amazon Mechanical Turk. *Journal of Second Language Pronunciation, 5*(2), 294–323. https://doi.org/10.1075/jslp.18016.nag

Noe-Bustamante, L., Flores, A., & Shah, S. (2019). Facts on Hispanics of Mexican Origin in the United States, 2017. *Pew Research Center.* https://www.pewresearch.org/hispanic/fact-sheet/u-s-hispanics-facts-on-mexican-origin-latinos/

Olson, D. J. (2022a). The Bilingual Code-switching Profile: Assessing the reliability and validity of the BCSP questionnaire. *Linguistic Approaches to Bilingualism*. Advance online publication. https://doi.org/10.1075/lab.21039.ols

Olson, D. J. (2022b). *A systematic review of proficiency assessment methods in bilingualism research.* [Unpublished manuscript].

Ortega-Santos, I. (2019). Crowdsourcing for Hispanic linguistics: Amazon's mechanical turk as a source of Spanish data. *Borealis: An International Journal of Hispanic Linguistics, 8*(1), 187–215. https://doi.org/10.7557/1.8.1.4670

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavioral Research Methods*, *46*(4), 1023–1031. https://doi.org/10.3758/s13428-013-0434-y

Perpiñán, S. (2018). On convergence, ongoing language change, and crosslinguistic influence in direct object expression in Catalan–Spanish bilingualism. *Languages*, *3*(2), 14. https://doi.org/10.3390/languages3020014

Poarch, G. J., Vanhove, J., & Berthele, R. (2019). The effect of bidialectalism on executive function. *International Journal of Bilingualism*, *23*(2), 612–628. https://doi.org/10.1177/1367006918763132

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org

Rahman, A., Yap, N. T., & Darmi, R. (2018). The association between vocabulary size and language dominance of bilingual Malay-English undergraduates. *3L, Language, Linguistics, Literature*, *24*(4). https://doi.org/10.17576/3L-2018-2404-07

Romaine, S. (1999). *Bilingualism* (2nd ed.). Wiley.

Satraki, M. (2019). Language attitudes: An overview. *International Journal of Linguistics, Literature, and Translation, 2*(6), 98–113.

Schleef, E. (2022). Measuring language attitudes. In K. Geeslin (Ed.), *The Routledge handbook of second language acquisition and sociolinguistics* (pp. 212–223). Routledge.

Solís-Barroso, C., & Stefanich, S. (2019). Measuring language dominance in early Spanish/English bilinguals. *Languages*, *4*(3), 62. https://doi.org/10.3390/languages4030062

Stocker, L., & Berthele, R. (2020). The roles of language mode and dominance in French–German bilinguals' motion event descriptions. *Bilingualism: Language and Cognition*, *23*(3), 519–531. https://doi.org/10.1017/S1366728919000294

Thomas, M. (1994). Assessment of L2 proficiency in second language acquisition research. *Language Learning*, *44*, 307–307. https://doi.org/10.1111/j.1467-1770.1994.tb01104.x

Tomé Lourido, G. (2018). *The role of social factors in bilingual speech processing: The case of Galician New Speakers*. [Doctoral dissertation, University College London]. https://discovery.ucl.ac.uk/id/eprint/10041801/

Tomoschuk, B., Ferreira, V. S., & Gollan, T. H. (2019). When a seven is not a seven: Self-ratings of bilingual language proficiency differ between and within language populations. *Bilingualism: Language and Cognition, 22*(3), 516–536. https://doi.org/10.1017/S1366728918000421

Treffers-Daller, J. (2011). Operationalizing and measuring language dominance. *International Journal of Bilingualism*, *15*(2), 147–163. https://doi.org/10.1177/1367006910381186

Treffers-Daller, J. (2016). Language dominance: The construct, its measure, and operationalization. In C. Silva-Corvalán & J. Treffers-Daller (Eds.), *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (pp. 235–265). Cambridge University Press. https://doi.org/10.1017/CBO9781107375345.005

Treffers-Daller, J. (2019). What defines language dominance in bilinguals?. *Annual Review of Linguistics*, *5*, 375–393. https://doi.org/10.1146/annurev-linguistics-011817-045554

Unsworth, S. (2016). Amount of exposure as a proxy for dominance in bilingual language acquisition. In C. Silva-Corvalán and J. Treffers-Daller (Eds.), *Language Dominance in Bilinguals: Issues of Measurement and Operationalization* (p. 156–173). Cambridge University Press. https://doi.org/10.1017/CBO9781107375345.008

Unsworth, S., Chondrogianni, V., & Skarabela, B. (2018). Experiential measures can be used as a proxy for language dominance in bilingual language acquisition research. *Frontiers in Psychology*, 1809. https://doi.org/10.3389/fpsyg.2018.01809

U.S. Census. (2020). *Dashboard – United States: Hispanic or Latino percent.* https://www.census.gov/quickfacts/fact/dashboard/US/RHI725219

van der Broek, E. W., Oolbekkink-Marchand, H. W., van Kemenade, A. M., Meijer, P. C., & Unsworth, S. (2022). Stimulating language awareness in the foreign language classroom:

exploring EFL teaching practices. *The Language Learning Journal*, *50*(1), 59–73. ! https://doi.org/10.1080/09571736.2019.1688857 !

Wei, L. (2000). Dimensions of bilingualism. In L. Wei (Ed.), *The bilingualism reader* (pp. 3–25). Routledge.

Yip, V., & Matthews, S. (2006). Assessing language dominance in bilingual acquisition: A case for mean length utterance differentials. *Language Assessment Quarterly: An International Journal*, *3*(2), 97–116. https://doi.org/10.1207/s15434311laq0302_2