Journal of the International Neuropsychological Society (2022), **28**, 845–861
Copyright © INS. Published by Cambridge University Press, 2021.
doi:10.1017/S1355617721001004

The MINT Sprint: Exploring a Fast Administration Procedure with an Expanded Multilingual Naming Test

Dalia L. Garcia[1,*] and Tamar H. Gollan[2,*]
[1]Joint Doctoral Program in Language and Communicative Disorders, San Diego State University/University of California, San Diego, CA, USA
[2]Department of Psychiatry, University of California, San Diego, CA, USA

(Received November 3, 2020; Final revision May 19, 2021; Accepted June 23, 2021; First Published Online August 31, 2021)

Abstract

Objectives: The present study examined if time-pressured administration of an expanded Multilingual Naming Test (MINT) would improve or compromise assessment of bilingual language proficiency and language dominance. **Methods:** Eighty Spanish–English bilinguals viewed a grid with 80 MINT-Sprint pictures and were asked to name as many pictures as possible in 3 min in each language in counterbalanced order. An Oral Proficiency Interview rated by four native Spanish–English bilinguals provided independent assessment of proficiency level. Bilinguals also self-rated their proficiency, completed two subtests of the Woodcock-Muñoz, and a speeded translation recognition test. We compared scores after 2 min, a first-pass through all the pictures, and a second-pass in which bilinguals were prompted to try to name skipped items. **Results:** The MINT Sprint and a subset score including original MINT items were highly correlated with Oral Proficiency Interview scores for predicting the degree of language dominance – matching or outperforming all other measures. Self-ratings provided weaker measures (especially of degree of balance – i.e., bilingual index scores) and did not explain any unique variance in measuring the degree of language dominance when considered together with second-pass naming scores. The 2-min scoring procedure did not improve and appeared not to hamper assessment of absolute proficiency level but prompting to try to name skipped items improved assessment of language dominance and naming scores, especially in the nondominant language. **Conclusions:** Time-pressured rapid naming saves time without significantly compromising assessment of proficiency level. However, breadth of vocabulary knowledge may be as important as retrieval speed for maximizing the accuracy in proficiency assessment.

Keywords: Bilingualism, Language dominance, Self-ratings, Speeded naming, Oral Proficiency Interview (OPI)

INTRODUCTION

Language proficiency is highly sought after in clinical, professional, and educational settings. In clinical settings, accurately measuring the language proficiency is critical for making accurate diagnoses (Bedore & Peña, 2008; Bedore et al., 2012; Gasquoine & Gonzalez, 2012), but it is common practice to simply ask the patients which language they prefer or to test in the majority language regardless of proficiency level. While self-ratings are an easy way to obtain some information about proficiency, self-ratings rely on participants' perception of their own linguistic abilities, which are influenced by factors that introduce considerable noise. Despite this problem, reliance on self-ratings is common because bilingual psychometrists, research assistants, and speech-language pathologists who can administer objective tests in both languages are not always available. Only 6.5% of clinicians meet the American Speech-Language-Hearing Association's definition of a bilingual service provider, which itself relies on self-identification as having native or near-native proficiency in a second language (American Speech-Language-Hearing Association, 2020).

Self-Rated Proficiency

Bilinguals are often asked to rate proficiency on a numerical scale. While self-ratings are significantly correlated with objective proficiency measures, the correlations tend to be small to moderate in size (Marian et al., 2007). Bilinguals are somewhat better in identifying which language is dominant, but self-ratings of absolute proficiency level and degree of bilingualism are far less accurate (i.e., whether proficiency level in the two languages is similar or not; Bedore et al., 2012; Gollan et al., 2012). Questionnaires also vary in how proficiency level is described and in the range of numerical

*Correspondence and reprint requests to: Department of Cognitive Science, UC San Diego, 9500 Gilman Drive, Mail Code 0515, La Jolla, CA 92093-0515, USA. E-mails: dgarcia1852@sdsu.edu; tgollan@ucsd.edu

scales used (e.g., 5-point, 7-point, and 10-point scales are common), further limiting the utility of self-ratings for comparison across studies.

Self-ratings are especially problematic when the goal is to compare across bilinguals of different language combinations and even within the same language combination if bilinguals are dominant in different languages or have a different learning history. Lemhöfer and Broersma (2012) examined self-rated proficiency in Dutch–English and Korean–English bilinguals using the same rating scale measured against the same objective tests. They used median splits to classify participants into large versus small vocabulary-size groups based on the ability to translate, self-ratings, and accuracy in a written lexical decision test (the LexTALE). Only 88.2% of Dutch–English bilinguals and 55.2% of Korean–English bilinguals accurately classified themselves into the correct vocabulary groups based on their translation performance. Similarly, the two bilingual groups were matched for the ability to translate, but Korean–English bilinguals rated themselves as significantly less English-proficient than Dutch–English bilinguals. Participants also rated themselves lower if they first completed the proficiency tests (for similar testing order effects on self-ratings, see Delgado et al., 1999).

Such between-group differences may reflect cultural or demographic differences in how rating scales are interpreted, reference scale, and/or standards of excellence (for related discussion see Nicoladis & Montanari, 2016; Rivera Mindt et al., 2010). Similar findings were reported in a study with self-ratings of 223 Chinese–English and 992 Spanish–English bilinguals tested across several studies (Tomoschuk et al., 2019) on the Multilingual Naming Test (MINT; Gollan et al. 2012; Sheng et al., 2014). Of bilinguals who gave themselves the maximum rating (7 on a 7-point scale), Chinese–English bilinguals correctly named 87% (59/68 MINT items) while Spanish–English bilinguals with the same self-rating averaged just 75% correct (51/68). It might seem that Chinese–English bilinguals are more accurate in their self-rating abilities; however, at the lower end of the scale, even larger discrepancies were found in the opposite direction. Chinese–English bilinguals who rated themselves a 3 (on the 7-point scale) averaged just 44% correct (30/68) in Chinese, while Spanish–English bilinguals with the same self-rating averaged 62% correct (42/68) in Spanish. Thus, across the two groups of bilinguals, the same ratings predicted different outcomes on the objective proficiency test in opposite directions at opposite ends of the rating scale, and within-group differences were found among speakers of the same languages but different dominance profiles. This makes it unlikely that any differences found simply reflected one group having better self-estimation abilities or that the test is easier in one language than the other. Self-ratings are also not comparable across different age groups. Older adults tend to rate their language abilities as being lower than young adults despite being matched on the ability to translate in both directions (Gollan et al., 2008; Gollan & Ferreira, 2009). Recent more elaborate approaches to self-assessment of bilingual language use (e.g., Gullifer & Titone, 2020) might be more accurate than simple self-ratings, but this possibility awaits further study.

Objective Proficiency Measures

While several studies demonstrated that objective proficiency tests are superior to self-ratings, there is no consensus as to which measures should be used and little information as to which measures work best for what purpose. One approach has been to use tests developed for English speakers in both languages, such as asking bilinguals to name pictures on the Boston Naming Test (BNT; Kaplan et al., 1983; Moreno et al., 2002; Silverberg & Samuel 2004). This is problematic because the test is often easier in the language for which it was developed. Gollan et al. (2012) found that the BNT characterized some relatively balanced bilinguals and even some Spanish-dominant bilinguals as English-dominant (see also Kohnert et al., 1998). Others designed tests with different items in each language; this is only better if difficulty is perfectly matched across languages – a substantial challenge (Peña, 2007).

Several studies used letter and semantic verbal fluency tasks to measure proficiency (Miranda et al., 2016; Rosselli et al., 2000; Zirnstein et al., 2018), and some have suggested that semantic fluency is especially “culturally fair” (Ardila & Moreno, 2001; Pekkala et al., 2009), while letter fluency is not (Artiola i Fortuny et al., 1998; Eng et al., 2019). However, fluency performance varies with specific categories (e.g., *animals* might be culture fair while *clothing* is not), and the fluency task does not measure proficiency alone, but also processing speed and executive control ability (e.g., application of strategies, switching, etc.). Thus, fluency tasks may be more affected by interference between languages and testing order than picture naming (Luo et al., 2010; Sandoval et al., 2010; Van Assche, et al., 2013) and may be more affected by idiosyncratic cross-linguistic differences (e.g., a language spoken in a tropical location might have more fruit names than a language spoken in the frozen tundra). Note however, that such idiosyncratic effects have also been identified in picture naming tests; heritage speakers who complete all their schooling in English may find it easier to name home items in Spanish (Bialystok et al., 2010; Wood et al., 2018).

Receptive vocabulary tests (e.g., the PPVT; Dunn & Dunn, 2007, and TVIP; Dunn, 1986; Umbel et al. 1992), especially written vocabulary tests, are convenient since they can be administered by experimenters who do not speak the languages. The LexTALE was developed to test proficiency level in English learners (Lemhöfer & Broersma, 2012) and was validated as a proficiency measure using a translation task and adapted to assess Spanish proficiency (LexTALE-Esp; Izura et al., 2014) following the same structure as the original LexTALE. However, ideally, objective measures should be developed in parallel for the two languages. Unlike the original LexTALE, the LexTALE-Esp was validated with self-ratings rather than with an independent

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

proficiency measure. Though self-ratings and LexTALE-Esp scores were correlated, Spanish learners who rated themselves a six or greater (on a 10-point scale) scored lower on the LexTALE-Esp than native Spanish speakers with the same self-ratings.

The Multilingual Naming Test (MINT)

The MINT was developed specifically to assess bilingual language proficiency (Gollan et al., 2012; Sheng et al., 2014) and was validated using Oral Proficiency Interviews (OPI), which provides a more comprehensive measure of language abilities including the ability to converse, express thoughts, and elaborate on complex ideas in production of full sentences. The MINT has 68 pictures arranged by difficulty level for both languages. Bilinguals name the same pictures in each of their two languages. This unique aspect of the MINT eliminates a source of noise that is introduced when bilinguals are asked to name different objects in each language in which lack of familiarity with one object will have an idiosyncratic effect on just one language.

The Present Study

The current study examined the potential utility of a time-pressured administration procedure. Rapid naming might improve proficiency assessment if the ability to retrieve names quickly forms a critical part of "language proficiency" as a construct. In psycholinguistic research, timed naming responses dominate as the measure of choice. In clinical settings, accuracy is typically measured, but it would be of great practical interest if proficiency could be assessed accurately under time restrictions. Alternatively, time-pressured administration could negatively affect proficiency assessment if untimed responses provide a better estimate of the size/breadth of the lexicon and if this is more closely tied to proficiency than naming speed.

In addition to the change in administration procedure in the MINT Sprint, we added a small number of more-difficult-to-name pictures, replaced black-and-white line drawings with colored pictures, and validated naming scores against OPI ratings provided by four independent raters to increase external validity (the original MINT had just one rater). The addition of more difficult items could improve proficiency assessment (especially in the dominant language), but was motivated by findings of ceiling effects in highly educated monolinguals (Stasenko et al., 2019). Two additional goals were to compare self-ratings of spoken proficiency to the average rating for all four modalities (speaking, listening, reading, and writing) and to compare the MINT Sprint to another timed test previously shown to be sensitive to proficiency level, a translation recognition test (Talamas et al., 1999). For additional comparison, bilinguals completed two subtests of another commonly used proficiency test (the Woodcock-Muñoz Language Survey; Woodcock et al., 2005).

METHODS

Participants

Eighty-one Spanish–English bilingual (64 female) undergraduates at the University of California, San Diego (UCSD), received course credit for participating. One was excluded for having incomplete data. Participant characteristics are shown in Table 1 with bilinguals divided into English-dominant ($n = 52$), balanced ($n = 25$), and Spanish-dominant ($n = 3$) groups based on their OPI scores. To classify bilinguals into these groups, we calculated dominance scores for each participant by subtracting the nondominant language score from the dominant language score. We calculated an average dominance score and standard deviation for all 80 bilinguals ($M = .09$; $SD = .10$).[1] Following Gollan et al., (2012) participants with dominance scores within half a standard deviation from zero (range −.04 to +.04) were classified as balanced, with those with positive scores above .04 were classified as English-dominant (range +.05 to +.41), and those with negative scores less than −.04 were classified Spanish-dominant (range −.05 to +.16). Note that English-dominant and balanced bilinguals differed significantly in just one demographic variable reported in Table 1 (current use of English).

Materials & Procedure

The research protocol was approved by the UCSD Institutional Review Board in accordance with the Helsinki Declaration. Participants signed a consent form and completed a Language History Questionnaire followed by the MINT Sprint and OPI in counterbalanced order in one language, followed by the Translation Recognition Task (TRT), and then the MINT Sprint and OPI in the other language. The Woodcock-Muñoz subtests were administered at the end in counterbalanced order beginning with the language most recently used and followed by the other language. Table 2 shows item characteristics for picture naming tests and the TRT. Table 3 shows performance on all tasks for each proficiency group.

Oral proficiency interview (OPI)

The OPI was designed based on the format used by the American Council on the Teaching of Foreign Languages and modified from Gollan et al., (2012) to update current events questions. Participants answered five questions beginning with easy "warm up" questions and then progressing to difficult questions designed to elicit higher-level language skills (e.g., complex sentence structures, defending an opinion). Participants also described a picture depicting a complex scene in each language.

Participants were interviewed by one of two proficient native Spanish–English bilingual experimenters, who both later listened to the recordings of all 80 interviews along with

[1] In Gollan et al. (2012), participants were placed into dominance groups (Spanish dominant, English dominant, or balanced) based on self-ratings. But since self-ratings are not reliable (Gollan et al., 2012; Lemhöfer & Broersma, 2012; Tomoschuk et al., 2019), we based dominance groups on the OPI.

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

Table 1. Participant characteristics from language history questionnaire

	English-dominant (*n* = 52)		Balanced (*n* = 25)		Spanish-dominant (*n* = 3)	
	M	*SD*	*M*	*SD*	*M*	*SD*
Age (in years)	20.3	1.6	20.1	2.7	18.7	1.2
Female	78.8	n/a	80.0	n/a	100.0	n/a
Education (in years)	13.8	1.1	13.5	1.8	12.6	1.2
Age First exposure to English	3.3	2.8	3.0	2.0	6.7	3.1
Age began using English regularly	4.6	2.8	4.7	3.3	8.7	4.2
Age First exposure to Spanish	.4	1.0	.2	.7	.7	.6
Age began using Spanish regularly	2.2	1.9	2.4	3.1	1.3	.6
Current percent of English use*	77.3	14.0	84.2	12.3	76.7	5.8
Percent of English use when growing up	51.4	14.3	56.6	17.8	30.0	20.0
Current percent of Spanish use	22.3	14.2	16.8	12.1	23.3	5.6
Percent of Spanish use when growing up	47.3	15.1	49.3	19.2	43.3	5.8
Language broker growing up[a]	55.8	n/a	48.0	n/a	66.7	n/a
How often codes switch currently[b]	3.9	1.4	3.9	1.5	5.7	.6
How often codes switch growing up[b]	3.8	1.5	3.6	1.4	4.3	2.1
Primary caregiver education (in years)	11.3	3.6	11.0	3.9	8.0	2.8
Secondary caregiver education (in years)	10.5	4.3	11.0	3.1	12.0	.0
			Self-Ratings of Proficiency[c]			
English speaking	6.6	.6	6.6	.9	6.3	.6
English reading	6.6	.6	6.5	.9	6.7	.6
English writing	6.4	.8	6.5	.8	6.0	1.0
English understanding	6.7	.6	6.6	.8	6.0	1.0
English average	6.6	.6	6.6	.8	6.3	.7
Spanish speaking	6.1	1.0	6.0	1.4	6.0	1.0
Spanish reading	5.8	1.0	5.9	1.2	6.0	.0
Spanish writing	5.5	1.1	5.6	1.1	5.3	.6
Spanish understanding	6.4	.9	6.2	1.4	6.3	1.2
Spanish average	5.9	.9	5.9	1.1	5.9	.6

[a] Reflects the percentage of participants that identified as language brokers.
[b] The following six-point scale was used: 1 = never or almost never; 2 = rarely, 3 = occasionally, 4 = two or three times in each conversation, 5 = several times in each conversation, 6 = a lot or sometimes even constantly.
[c] The following seven-point scale was used: 1 = almost none, 2 = very poor, 3 = fair, 4 = functional, 5 = good, 6 = very good, 7 = like a native speaker.
*Significant t-test comparing Balanced bilinguals to English-dominant bilinguals (*p* < .05).

two additional proficient Spanish–English bilingual raters. Final OPI scores were the average of ratings assigned by the four raters in each language on a 10-point scale with detailed scoring criteria (see Appendix A for OPI questions and scoring criteria).

MINT sprint

A set of 80 pictures were presented in an eight-by-ten grid simultaneously on a computer monitor. Items included colored pictures depicting all of the same objects in the original MINT in addition to a small number of more difficult items drawn from studies designed to elicit tip-of-the-tongue states (Gollan & Brown, 2006; Stasenko & Gollan, 2019). Appendix B presents a complete list of the MINT Sprint items. Three of new pictures had cognate names (which are formally similar across languages, e.g., *gyroscope* is *giróscopo* in Spanish); the remaining 77 were noncognates. Note that cognate status affects naming only when bilinguals know the word in both languages, which is increasingly unlikely for objects with very low frequency names (Gollan & Acenas, 2004). Using existing data from previous experiments in the lab, items were ordered by difficulty collapsing across both languages, with the easier items appearing in the top rows and the harder items at the bottom (see Appendix B for MINT Sprint items). To give participants a sense of time pressure, they were told they had 3 minutes to name as many pictures as they could, as quickly as possible starting at the top left corner and make their way across each row, and with permission to go back to name items they previously skipped (and without requirement to point to items as they named them). Most participants required less than 3 minutes for each language to complete their first-pass (first attempt) through the entire grid, and the 3 minutes cutoff was not imposed (participants were given as much time as they needed). After participants said they were finished, they were prompted to take a second-pass through all the pictures to try to name all the items they skipped in the first-pass.

The MINT Sprint testing materials and data from the current study have been uploaded to Open Science Framework. You can access these materials here.

Table 2. Item characteristics for translation recognition and picture naming tasks, frequency per million. Frequencies per million were calculated based on the SUBTLEX-US (Brysbaert & New, 2009) and the SUBTLEX-ESP (Cuetos et al., 2011)

	English								Spanish							
	Frequency[a,b]		Length in Syllables		Length in Letters		Length in Phonemes		Frequency[a,b]		Length in Syllables		Length in Letters		Length in Phonemes	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Translation Recognition Test	60.8	164.6	1.2	.4	5.0	1.4	4.0	1.2	120.4	209.4	2.3	.5	5.4	1.2	5.2	2.3
Original MINT Items	50.1	84.4	1.4	.5	5.1	1.7	3.8	1.3	36.4	62.4	2.6	.8	6.1	1.8	5.8	1.8
MINT Sprint Items	41.9	79.1	1.5	.7	5.5	1.9	4.2	1.6	30.9	58.6	2.6	.8	6.3	1.8	6.0	1.8
Woodcock-Muñoz Picture Vocabulary	28.5	78.6	2.2	1.0	6.8	2.2	5.7	2.3	37.6	83.9	2.9	1.0	6.8	2.3	6.6	2.4

[a] The following MINT Sprint Spanish items were not found in the SUBTLEX database: *pavo real, arco iris, yunque,* and *mentrómono.* The English item *mortar* was not found in the SUBTLEX. Missing items were not included in these descriptive statistics. The item *well* was excluded in English since the SUBTLEX database did not give specific information about frequency dependent on part of speech. The use of *well* as an adjective and adverb in English artificially inflates the count of the word.

[b] The following Spanish items from the Woodcock-Muñoz Picture Vocabulary subtest were not found in the SUBTLEX database: *San Basilio, barquilla, petroglifo,* and *plinto.* The English item *sabot* was not found in the SUBTLEX. Missing items were not included in these descriptive statistics. Also, for items that required a multiword response (*fire extinguisher, flamenco dancer, and filing cabinet*) only the frequency of the main noun was counted (*extinguisher, dancer, and cabinet*).

Translation recognition test (TRT)

In the TRT (Talamas et al., 1999), participants saw a Spanish word followed by an English word and were asked to decide if they are translation equivalents (a "yes" decision) or not (a "no" decision). There were 160 trials; half were translation equivalents, and half were evenly divided into four different types: semantically related to the translation (e.g. *jabón-bath, jabón* means *soap*), semantically unrelated matched control (e.g. *self*), translation form related (e.g., *soup*), and translation form unrelated matched control (e.g. *clay*). Talamas et al. (1999), reported "no" decisions were sensitive to proficiency; at low proficiency levels, bilinguals had more difficulty rejecting form-related than semantically related distractors, while sensitivity to semantically related distractors increased with increasing proficiency level. All stimuli were nouns, many taken from Ma et al. (2017) with some replacements to accommodate regional variations. Stimuli for the unrelated conditions were matched on length and frequency to the corresponding related condition.

Self-rated proficiency

Participants were presented with a Language History Questionnaire in which they rated their language abilities in four modalities (speaking, reading, understanding, writing) on a 7-point scale (1 – almost none, 7 – like a native speaker).

Woodcock-Muñoz picture vocabulary (WMPV)

Participants attempted to name all 47 pictures in the Spanish version and all 45 pictures in the English version of the WMPV subtest. Participants were presented with six pictures at a time and were asked to point to each item as they named it. Responses were scored in accordance to the Woodcock-Muñoz guide. If participants produced an answer that required further elicitation, the experimenter would prompt the participant to produce a different name for the picture.

Woodcock-Muñoz passage comprehension (WMPC)

Participants started both the English and Spanish Passage Comprehension subtests at the ninth-grade level (completing 22 sentences in Spanish and 20 in English). They were presented with four sentences at a time via a paper packet. Each sentence had a blank, and participants were instructed to produce a word that would fit in the blank (e.g. *Reptile eggs look a lot like bird eggs. Some are almost perfectly ____ like ping-pong balls; other are oblong* or *En la mayoría de las ___ hay muchos edificios altos*). If participants produced an answer that required further elicitation, they received a prompt to produce a different word.

RESULTS

For all proficiency measures, we examined correlations with OPI scores. In addition to MINT Sprint scores, we calculated

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

Table 3. Performance across tasks for bilinguals with different dominance profiles. Proportion correct or proportion of maximum possible for each task (e.g., 9/10 = .9 for oral proficiency interview)

	English-dominant ($n = 52$)		Balanced ($n = 25$)		Spanish-dominant ($n = 3$)	
	M	*SD*	*M*	*SD*	*M*	*SD*
English Oral Proficiency Interview	.913	.032	.906	.046	.863	.075
Spanish Oral Proficiency Interview	.769	.075	.892	.041	.958	.026
English Original MINT	.880	.058	.844	.083	.736	.129
Spanish Original MINT	.647	.116	.771	.094	.886	.071
English MINT Sprint Second Pass	.815	.070	.767	.101	.642	.125
Spanish MINT Sprint Second Pass	.577	.111	.700	.096	.833	.062
English Woodcock-Muñoz Picture Naming	.652	.088	.626	.113	.519	.100
Spanish Woodcock-Muñoz Picture Naming	.548	.084	.638	.056	.738	.033
English Woodcock-Muñoz Passage Comprehension	.459	.136	.498	.127	.417	.153
Spanish Woodcock-Muñoz Passage Comprehension	.435	.144	.545	.142	.636	.091
Translation Recognition Task	.925	.031	.944	.020	.927	.010

Table 4. Correlations between different proficiency measures and the oral proficiency interview scores for different language abilities[a] (n = 80)

	MINT Sprint 2-min scoring procedure	MINT Sprint First-Pass	MINT Sprint Second-Pass Score	Original Mint Items	Woodcock-Muñoz Picture Vocabulary	Woodcock-Muñoz Sentence Completion	Average Self-Rating	Spoken Self-Rating[c]	Translation Recognition Overall Accuracy
English	.455	.455	.451	.431	.469	.235[b]	.447	.462	.017[d]
Spanish	.589	.601	.662	.659	.650	.576	.421	.394	.318
Dominance	.733	.740	.780	.773	.740	.435	.569	.537	−.293
Index	.634	.636	.691	.694	.455	.287	.393	.298	.323

[a] Unless otherwise indicated all correlations significant at $p < .001$.
[b] $p < .05$.
[c] *n = 79, one bilingual left the self-rating for spoken proficiency in Spanish blank.*
[d] not significant.

an *original MINT* score for each bilingual based on the original 68 MINT items (but note that color pictures replaced the original black-and-white line drawings). Our greatest interest was the magnitude of correlations between OPI scores and the MINT Sprint, the subset of original MINT items, spoken self-ratings, and self-ratings averaged across the four modalities using Steiger's Z-test after applying Fischer's r-to-Z transformations (to normalize the distribution of *r*-values). The correlations with OPI scores are shown in Table 4, and correlations among all measures are shown in Appendix C. Scores for all tasks were converted to proportions for comparability (i.e., proportion correct for naming scores and proportion of total possible score for self-ratings and OPI e.g., 9/10 = 90% or .9). For the TRT, we report overall accuracy (the number of correct yes and no responses divided by the number of trials) because this task does not provide separate scores for each language and of all the possible TRT measures (e.g., response times, distractor conditions; see Appendix D), and overall accuracy was the only measure that was significantly correlated with OPI scores. None of the tasks exhibited language of testing order effects (all *p*s ≥ .15).

English

English OPI scores ranged from 7.75 to 9.88 (out of 10; see Appendix A). Most participants were English-dominant (see Table 1) and scored near ceiling on the OPI (OPI; $M = 9.09$, $SD = .40$). Picture naming tests, self-rated proficiency, and WMPC were equally correlated with English OPI scores (all *r*s between .24 and .47; none of these differed from each other using Steiger's Z-test, all *p*s ≥ .11). Performance on the TRT was not correlated with English OPI scores ($r = .02$, $p = .88$).

Spanish

Spanish OPI scores ranged from 5.50 to 9.88. Picture naming tests were the best predictors of Spanish OPI scores; (both *r*s between .65 and .66). The WMPC, TRT, and self-ratings were only moderately correlated with OPI scores (*r*s between .32 and .42). The MINT Sprint and original MINT score were significantly more correlated with OPI scores than self-ratings of spoken proficiency and self-ratings averaged across all four modalities (all *z*s between 2.01 and 2.24, all *p*s < .05).

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

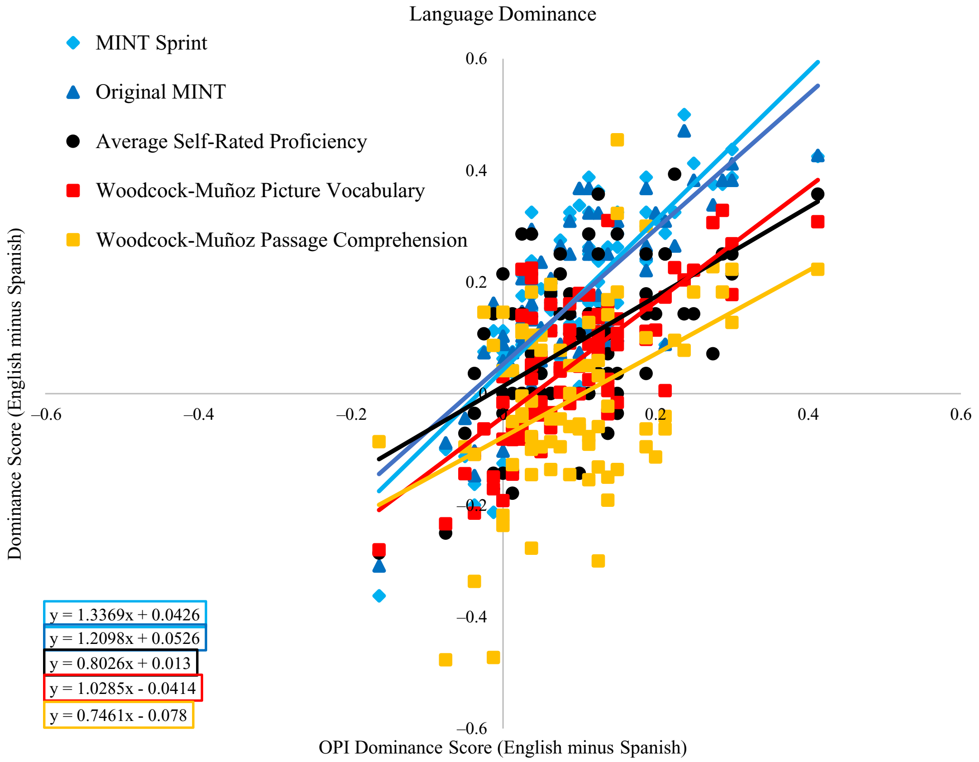

Fig. 1. The x-axis shows language dominance scores (English minus Spanish) on the Oral Proficiency Interview (OPI). Negative scores along both axes show bilinguals who scored higher in Spanish than in English.

Language Dominance

We calculated a language dominance score for each measure by subtracting proportion correct in Spanish from proportion correct in English. The picture naming tests were correlated with OPI dominance scores at higher than $r = .70$, while the WMPC subtest and self-ratings correlations ranged between .43 and .57. The TRT was negatively correlated with OPI scores ($r = -.29$) indicating that more English-dominant bilinguals had greater difficulty recognizing translation equivalents. The MINT Sprint and original MINT score were significantly more correlated with OPI scores than both spoken self-ratings and average self-ratings (zs between 2.28 and 2.65, *ps* < .05).

Figure 1 shows relationships between the OPI scores and the other measures. The MINT Sprint and original MINT score fared best for predicting OPI scores (the lines with the steepest slope and y-intercept close to zero), although both exhibited some bias toward English dominance (y-intercept above 0). The WMPC subtest exhibited bias toward Spanish, classifying many bilinguals as Spanish-dominant (negative on the y-axis) who were classified as English-dominant (positive on the x-axis) on the OPI (this line has the y-intercept that is farthest away from zero relative to the other lines).

Bilingual Index

Bilingual index scores were calculated by dividing the language with the lower score by the language with the higher score (Gollan et al., 2012). For example, a bilingual who named 45 pictures in Spanish and 60 in English would have a bilingual index score of .75 (45/60) as would someone who named 45 in English and 60 in Spanish. Thus, index scores reflect the degree of balanced knowledge of the two languages while ignoring which language is dominant. The original MINT and the MINT Sprint again exhibited the highest correlations (approaching .70), while the other measures ranged between .29 and .45. Both the original MINT and the MINT Sprint were significantly more correlated with the OPI than both spoken and average self-ratings (all *zs* between 2.54 and 3.25, all *ps* < .05).

Figure 2 shows relationships between the OPI scores and the other measures. The MINT Sprint and original MINT score rated bilinguals as less balanced than the OPI but did so consistently for unbalanced and balanced bilinguals alike, thereby better preserving the rank order of the OPI index scores than all other measures (i.e., the blue lines are the steepest, with a slope closest to 1). By contrast, the other measures classified unbalanced bilinguals as *more* balanced than they were (especially self-ratings, the black line), but classified balanced bilinguals as *less* balanced than they were (especially the WMPC, the yellow line).

Self-rated proficiency

Table 4 shows that average self-ratings tended to fare slightly better in predicting OPI scores than self-ratings of just spoken proficiency, but these differences were not significant (all zs ≥ .65, all *ps* ≤ .52). To determine if average self-ratings captured any unique variance in predicting OPI scores, we

Table 5. Stepwise regression models with second-pass MINT sprint scores and average self-rated proficiency entered as independent variables and the oral proficiency interview (OPI) in English, Spanish, OPI dominance scores, and the OPI bilingual index scores as the dependent variables

Dependent variable	step		R^2	ΔR^2	ΔF
English	1	MINT Sprint	.204	.204	19.94***
	2	Average Self-Rated Proficiency	.274	.071	7.49**
Spanish	1	MINT Sprint	.442	.442	60.97***
	2	Average Self-Rated Proficiency	.471	.029	4.14*
Dominance	1	MINT Sprint	.617	.617	124.10***
	2	Average Self-Rated Proficiency	.624	.007	1.41
Index	1	MINT Sprint	.489	.489	73.58***
	2	Average Self-Rated Proficiency	.518	.029	4.62*

*$p < .05$, **$p < .01$, ***$p < .001$.

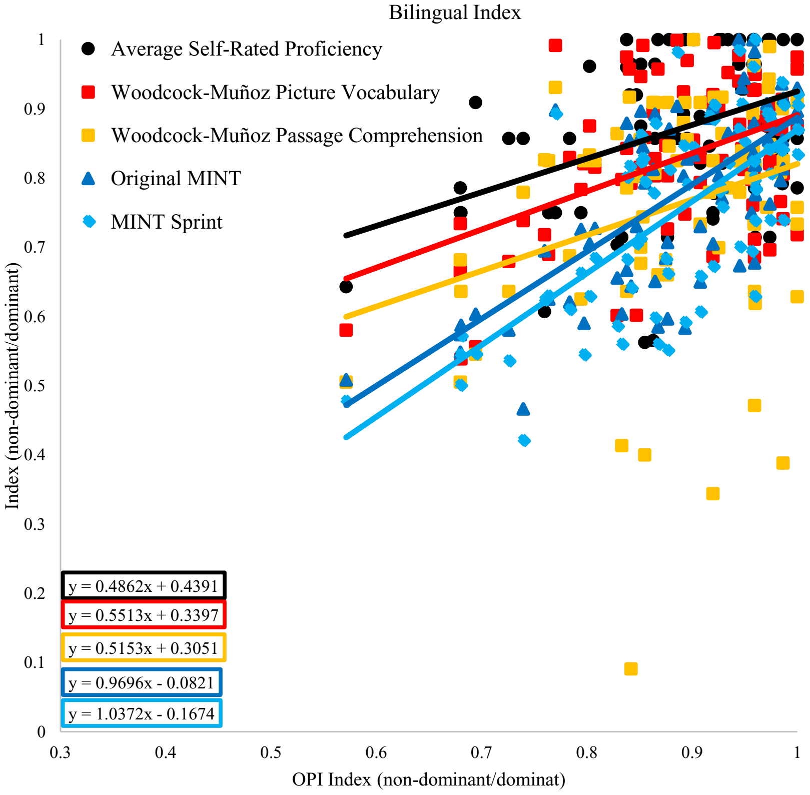

Fig. 2. The *x*-axis shows Bilingual Index Scores (Lower Score/Higher Score) on the Oral Proficiency Interview (OPI). This *y*-axis shows relationships between index scores across the different measures in comparison to the OPI index scores.

conducted stepwise linear regressions with OPI scores as the dependent variable and second-pass MINT Sprint scores alone versus with average self-rated proficiency added as predictors (see Table 5). All regression models were significant, and the MINT Sprint accounted for about 40%–50% of the variance for Spanish, language dominance, and bilingual index scores and about 20% for English. When adding self-ratings, these explained between 1% and 7% of additional variance in OPI scores for English, Spanish, and the bilingual index. However, for language dominance scores, average self-ratings did not explain any unique variance. Stepwise regressions with both independent variables were significantly better than the simple regressions for predicting English, Spanish, and bilingual index scores, but not language dominance, but overall R^2 changes when adding self-rating scores were relatively small.

MINT Sprint timing

To examine the possible effects of imposing time limits on assessment of productive vocabulary, we correlated OPI scores with three different MINT naming scores (see Table 4) including the number of pictures named in each language after: (a) 2 minutes,[2] (b) bilinguals completed a first-pass through the grid, and (c) bilinguals were prompted to take a second-pass through the grid to try to name any items they had skipped on the first-pass. Correlations with OPI scores did not differ significantly across different scoring procedures for predicting English, Spanish, language dominance, or the bilingual index OPI scores (all $zs < .7$, all $ps \geq .51$). Asking the same question in a different way leads to different conclusions. After 2-min, bilinguals named $M = 62.2$ ($SD = 6.6$) pictures in the dominant language and just $M = 42.1$ ($SD = 9.7$) in the nondominant language, a difference of 20.1. When the same bilinguals completed the whole grid without time restriction, henceforth *first-pass scores* (dominant $M = 63.3$, $S = 6.1$; nondominant $M = 43.4$, $SD = 10.1$), the difference shrank only very slightly to 19.8. However, with prompting to go back and name skipped items, henceforth *second-pass scores* (dominant $M = 64.9$, $SD = 5.6$; nondominant $M = 48.6$, $SD = 8.8$), the difference shrank to 16.3. Comparing the 2-min and second-pass scores in an ANOVA with score type (2-min scoring procedure, second-pass) and language (dominant, nondominant) as repeated measures factors revealed higher scores for second-pass scores, a main effect of score type, $F(1,79) = 290.06$, $p < .001$, $\eta_p^2 = .786$, $MSE = .001$, higher scores in the dominant than the nondominant language, a main effect of language $F(1,79) = 299.52$, $p < .001$, $\eta_p^2 = .791$, $MSE = .014$, and the nondominant language benefitted more from the second-pass prompt, a significant interaction between score type and language $F(1,79) = 50.35$, $p < .001$, $\eta_p^2 = .389$, $MSE = .001$; see Figure 3. Repeating the same analysis but comparing first-pass (instead of the 2-min scoring procedure) to second-pass scores revealed similar results (all $ps < .001$).

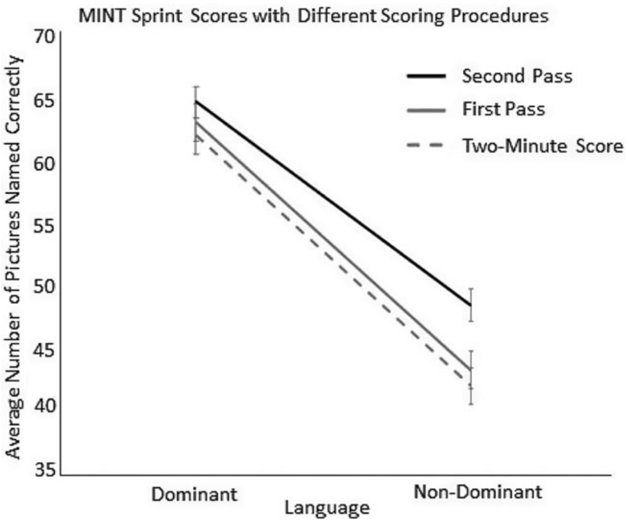

Fig. 3. Average MINT Sprint Scores in 2-min scoring procedure, first-pass, and second-pass scoring for all 80 bilinguals out of 80 MINT Sprint pictures in each language. Error bars show standard errors.

DISCUSSION

Summarizing the results, tests of picture naming fared best for predicting proficiency in Spanish (the nondominant language for most participants), language dominance, and degree of bilingualism (the bilingual index scores). By contrast, self-ratings and the other objective tests (WMPC and the TRT) did not fare as well, with correlations tending to be low or moderate at best. While self-ratings were not as good as picture naming for estimating oral proficiency scores (see also Gollan et al., 2012; Lemhöfer & Broersma, 2012; Sheng et al., 2014; Tomoschuk et al., 2019), self-ratings were relatively better at predicting language dominance than they were at predicting absolute proficiency level (in English and Spanish) and bilingual index scores (see Table 4). Additionally, self-ratings did explain small amounts of unique variance when combined with MINT Sprint scores for predicting English, Spanish, and bilingual index scores but not language dominance scores (see Table 5). Finally, imposing time limits (the time-saving administration procedure and the 2-min scoring procedure) did not seem to improve or compromise the utility of the MINT for rank-ordering bilinguals with respect to language proficiency. However, prompting bilinguals to go back and try to name items they skipped initially benefitted naming scores in the nondominant language more than the dominant language. These results highlight the importance of using objective measures especially to rank order bilinguals for degree of language dominance and degree of bilingualism (Gollan et al., 2012; Izura et al., 2014; Lemhöfer & Broersma, 2012; Tomoschuk et al., 2019), and have implications for understanding bilingual langauge proficiency, as follows:

A priori we anticipated that the pressure to name as many pictures as possible in a short amount of time could be useful in clinical settings (to ensure bilinguals are tested mostly in the language that will maximize performance) and might improve proficiency assessment by tapping retrieval speed, or alternatively, that the speeded component would come at a cost of assessing the breadth of lexical knowledge. Though second-pass scores tended to exhibit higher correlations with OPI scores than first-pass and 2-min scores (see Table 4), these were statistically equivalent, which could imply a trade-off between retrieval speed and breadth of lexical knowledge – so that both may be equally important. However, the conclusion 2-min was as good as second-pass scores for estimating proficiency is based on a null effect. We caution against interpreting the null because second-pass scores always tended in the direction of stronger correlation

[2] Pilot data suggested that instructing participants to name all the pictures in three minutes would elicit a strategy of naming as many pictures as quickly as possible, while examining scores after two minutes would allow sufficient time to discriminate between participants of different proficiency levels (whereas after just one minute only relatively easy pictures would be named by all). The 3-minute cutoff was not imposed (see Procedure) to make it possible to examine different possible scoring approaches.

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

with OPI scores, and the second-pass prompt improved nondominant more than dominant language scores (see Figure 3). Thus, assessment with limited time might overestimate the degree to which one language is dominant over the other.

Additional evidence suggesting that breadth of lexical knowledge is more critical than processing speed was found in the results of the TRT. Though participants exhibited robust condition effects in this task (see Appendix D), accuracy was the only measure that was significantly correlated with OPI scores. If processing speed was as important as breadth of lexical knowledge, we should have seen significant correlations in these data. However, the TRT items may not have been difficult enough to be sensitive to differences in proficiency level (accuracy was close to ceiling on this task; see Table 4). Additionally, having difficult items may be necessary but not sufficient; for reasons that might be idiosyncratic to the WMPC task, accuracy was far lower on the WMPC, which did not fare well in predicting OPI scores (see Tables 3 and 4; Figures 1 and 2). What might be critical is having a range of item difficulty (see also Ivanova et al., 2013; Kohnert et al., 1998). Our procedure of starting at the ninth-grade level on the WMPC may have compromised its sensitivity (but note that others have drawn similar conclusions about the Woodcock-Muñoz; Miranda et al., 2016). Importantly, it is not likely that WMPC and TRT were less correlated with OPI scores simply because they measure language comprehension while both the OPI and the MINT assessed language production. Hoversten and Traxler (2020) used the original MINT, the LexTALE, and LexTALE-Esp to assess proficiency in Spanish–English bilinguals. Combining 116 participants from both experiments in that paper, lexical decision scores were significantly correlated with original MINT scores in English ($r = .40$, $p < .001$), Spanish ($r = .46$, $p < .001$), language dominance ($r = .63$, $p < 001$), and the bilingual index ($r = .56$, $p < .001$; Hoversten, personal communication). Thus, comprehension measures can produce significant correlations with oral-proficiency level, and it is possible these correlations would have been even higher if the two versions of the LexTALE were developed to be comparable in the two languages.

Limitations and future directions

The OPI exclusively captures spoken language proficiency while excluding other domains of competence (e.g., auditory comprehension, reading and writing skills). The OPI also relies on subjective proficiency ratings, which, though not objective, are likely to be ecologically valid and relatively better for rank-ordering individuals by proficiency level because each rater assessed all 80 bilinguals on the same scale. We improved on our approach in developing the original MINT by having four raters for each bilingual (instead of a single rater as in Gollan et al., 2012). While the MINT Sprint administration procedure seemed to work very well for bilinguals, the items will need to be validated with monolingual speakers of each language to determine if item difficulty is equivalent across languages. Like the original MINT, the MINT Sprint may be easier in English (see Figure 1). Furthermore, Spanish items tended to be longer (see

Table 2), potentially introducing constant noise in the comparison of time-pressured naming abilities across languages. Additionally, the time-pressured administration procedure will need to be validated with older bilinguals and patients; simultaneous presentation of all 80 items might lead such participants to skip more items initially, making the second-pass even more crucial. Finally, though it appeared not to improve assessment in bilinguals, it is possible that time-pressured naming and scoring procedures might improve assessment of naming ability relatively more in monolinguals (e.g., see Stiver et al., 2021), who tend to be hyper-proficient in the one language they speak, especially at higher education levels (Stasenko et al., 2019).

Conclusion

The MINT Sprint accurately measures proficiency, language dominance, and degree of bilingualism. While the rapid administration procedure saves time and is likely adequate for many purposes, allowing a bit more time and prompting a second attempt at missed items likely maximizes accuracy in assessment of naming ability in the nondominant language and the degree of language dominance. Although self-ratings did not improve rank ordering of bilinguals by degree of language dominance, this is not an open invitation to ignore bilinguals' stated preferences for one language over another. Self-ratings may capture aspects of competence and personal preferences that could affect performance on some abilities (not tested herein), and though they must be interpreted with caution, self-ratings should always be interpreted in concert with objective measures and also considering the goals of the assessment. That said, picture naming tests are superior for rank-ordering bilinguals in proficiency level; self-ratings were biased toward a truncated range (all bilinguals rated themselves as relatively balanced; see Figure 2), and some self-classifications of language dominance were also incorrect (see Gollan et al., 2012). Given that the MINT Sprint can be administered in both languages in relatively little time, we hope it will increase use of a more rigorous approach to assessing bilingual language proficiency in both clinical and research settings.

ACKNOWLEDGMENTS

The authors thank Mayra Murillo for data collection and coding, Rosa Montoya and Rosemary Vela for data coding, and Tamar Degani for constructive feedback on this manuscript.

FINANCIAL SUPPORT

This research was supported by grants from the National Institute on Deafness and Other Communication Disorders

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

(011492) and the National Science Foundation (BCS 1923065). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NIH or NSF.

CONFLICT OF INTEREST

The authors have nothing to disclose.

REFERENCES

American Speech-Language-Hearing Association. (2020). *Demographic profile of ASHA members providing bilingual services, year-end 2019*. Available from www.asha.org

Ardila, A. & Moreno, S. (2001). Neuropsychological test performance in Aruaco Indians: an exploratory study. *Journal of the International Neuropsychological Society: JINS*, *7*, 510–515. doi: 10.1017/s1355617701004076

Artiola i Fortuny, L., Heaton, R. K., & Hermosillo, D. (1998). Neuropsychological comparisons of Spanish-speaking participants from the U.S.-Mexico border region versus Spain. *Journal of the International Neuropsychological Society: JINS*, *4*, 363–379.

Bedore, D.L.M. & Peña, E.D. (2008). Assessment of Bilingual children for identification of language impairment: current findings and implications for practice. *International Journal of Bilingual Education and Bilingualism*, *11*, 1–29. doi: 10.2167/beb392.0

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*, 616–629. doi: 10.1017/S1366728912000090

Bialystok, E., Luk, G., Peets, K.F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism (Cambridge, England)*, *13*, 525–531. doi: 10.1017/S1366728909990423

Brysbaert, M. & New, B. (2009). Moving beyond Kučera and Francis: a critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*, 977–990. doi: 10.3758/BRM.41.4.977

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, *32*, 133–143.

Delgado, P., Guerrero, G., Goggin, J.P., & Ellis, B.B. (1999). Self-assessment of linguistic skills by Bilingual Hispanics: *Hispanic Journal of Behavioral Sciences*. doi: 10.1177/0739986399211003

Dunn, L.M. (1986). *Test de vocabulario en ima□?genes Peabody (TVIP)*. Minnesota: AGS.

Dunn, L.M., Dunn, D.M., & Pearson Assessments. (2007). *PPVT-4: Peabody picture vocabulary test*. Minneapolis, MN: Pearson Assessments.

Eng, N., Vonk, J.M., Salzberger, M., & Yoo, N. (2019). A cross-linguistic comparison of category and letter fluency: Mandarin and English. *Quarterly Journal of Experimental Psychology*, *72*, 651–660. doi: 10.1177/1747021818765997

Gasquoine, P.G. & Gonzalez, C.D. (2012). Using monolingual neuropsychological test norms with bilingual Hispanic Americans: application of an individual comparison standard. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *27*, 268–276. doi: 10.1093/arclin/acs004

Gollan, T.H. & Acenas, L.-A. R. (2004). What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and tagalog-English bilinguals. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*, 246–269. doi: 10.1037/0278-7393.30.1.246

Gollan, T.H. & Brown, A.S. (2006). From tip-of-the-tongue (TOT) data to theoretical implications in two steps: when more TOTs means better retrieval. *Journal of Experimental Psychology. General*, *135*, 462–483. doi: 10.1037/0096-3445.135.3.462

Gollan, T.H., & Ferreira, V.S. (2009). Should I Stay or Should I Switch? A cost–benefit analysis of voluntary language switching in young and aging Bilinguals. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*, 640–665. doi: 10.1037/a0014981

Gollan, T.H., Montoya, R.I., Cera, C., & Sandoval, T.C. (2008). More use almost always a means a smaller frequency effect: aging, bilingualism, and the weaker links hypothesis. *Journal of Memory and Language*, *58*, 787–814. doi: 10.1016/j.jml.2007.07.001

Gollan, T.H., Weissberger, G.H., Runnqvist, E., Montoya, R.I., & Cera, C.M. (2012). Self-ratings of Spoken Language Dominance: A Multi-Lingual Naming Test (MINT) and Preliminary Norms for Young and Aging Spanish-English Bilinguals. *Bilingualism (Cambridge, England)*, *15*, 594–615. doi: 10.1017/S1366728911000332

Gullifer, J.W. & Titone, D. (2020). Characterizing the social diversity of bilingualism using language entropy. *Bilingualism: Language and Cognition*, *23*, 283–294. doi: 10.1017/S1366728919000026

Hoversten, L.J. & Traxler, M.J. (2020). Zooming in on zooming out: partial selectivity and dynamic tuning of bilingual language control during reading. *Cognition*, *195*, 104118. doi: 10.1016/j.cognition.2019.104118

Ivanova, I., Salmon, D.P., & Gollan, T.H. (2013). The multilingual naming test in alzheimer's disease: clues to the origin of naming impairments. *Journal of the International Neuropsychological Society : JINS*, *19*, 272–283. doi: 10.1017/S1355617712001282

Izura, C., Cuetos, F., & Brysbaert, M. (2014). Lextale-Esp: a test to rapidly and efficiently assess the Spanish vocabulary size. *Psicológica*, *35*, 49–66.

Kaplan, E., Goodglass, H., & Weintraub, S. (1983). *The Boston Naming Test*. Philadelphia: Lea & Fibiger.

Kohnert, K.J., Hernandez, A.E., & Bates, E. (1998). Bilingual performance on the Boston naming test: preliminary norms in Spanish and English. *Brain and Language*, *65*, 422–440. doi: 10.1006/brln.1998.2001

Lemhöfer, K. & Broersma, M. (2012). Introducing LexTALE: a quick and valid Lexical test for advanced learners of English. *Behavior Research Methods*, *44*, 325–343. doi: 10.3758/s13428-011-0146-0

Luo, L., Luk, G., & Bialystok, E. (2010). Effect of language proficiency and executive control on verbal fluency performance in bilinguals. *Cognition*, *114*, 29–41. doi: 10.1016/j.cognition.2009.08.014

Ma, F., Chen, P., Guo, T., & Kroll, J.F. (2017). When late second language learners access the meaning of L2 words: Using ERPs to investigate the role of the L1 translation equivalent. *Journal of Neurolinguistics*, *41*, 50–69. doi: 10.1016/j.jneuroling.2016.09.006

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: JSLHR*, *50*, 940–967. doi: 10.1044/1092-4388(2007/067)

Miranda, C., Rentería, M.A., Fuentes, A., Coulehan, K., Arentoft, A., Byrd, D., Rosario, A., … Mindt, M.R. (2016). The relative utility of three english language dominance measures in predicting the neuropsychological performance of HIV+ Bilingual Latino/a adults. *The Clinical Neuropsychologist*, *30*, 185–200. doi: 10.1080/13854046.2016.1139185

Moreno, E.M., Federmeier, K.D., & Kutas, M. (2002). Switching languages, switching palabras (words): an electrophysiological study of code switching. *Brain and Language*, *80*, 188–207. doi: 10.1006/brln.2001.2588

Nicoladis, E. & Montanari, S. (2016). *Bilingualism Across the Lifespan: Factors Moderating Language Proficiency*. https://www.apa.org/pubs/books/4316172

Pekkala, S., Goral, M., Hyun, J., Obler, L.K., Erkinjuntti, T., & Albert, M.L. (2009). Semantic verbal fluency in two contrasting languages. *Clinical Linguistics & Phonetics*. doi: 10.1080/02699200902839800

Peña, E.D. (2007). Lost in translation: methodological considerations in cross-cultural research. *Child Development*, *78*, 1255–1264. doi: 10.1111/j.1467-8624.2007.01064.x

Rivera Mindt, M., Byrd, D., Saez, P., & Manly, J. (2010). Increasing culturally competent neuropsychological services for ethnic minority populations: a call to action. *The Clinical Neuropsychologist*, *24*, 429–453. doi: 10.1080/13854040903058960

Rosselli, M., Ardila, A., Araujo, K., Weekes, V.A., Caracciolo, V., Padilla, M., & Ostrosky-Solí, F. (2000). Verbal fluency and repetition skills in healthy older Spanish-English Bilinguals. *Applied Neuropsychology*, *7*, 17–24. doi: 10.1207/S15324826AN0701_3

Sandoval, T., Gollan, T., Ferreira, V., & Salmon, D. (2010). *What Causes the Bilingual Disadvantage in Verbal Fluency? : The Dual-task Analogy*. doi: 10.1017/S1366728909990514

Sheng, L., Lu, Y., & Gollan, T.H. (2014). Assessing language dominance in Mandarin-English bilinguals: convergence and divergence between subjective and objective measures. *Bilingualism (Cambridge, England)*, *17*, 364–383. doi: 10.1017/S1366728913000424

Silverberg, S. & Samuel, A.G. (2004). The effect of age of second language acquisition on the representation and processing of second language words. *Journal of Memory and Language*, *51*, 381–398. doi: 10.1016/j.jml.2004.05.003

Stasenko, A. & Gollan, T.H. (2019). Tip of the tongue after any language: reintroducing the notion of blocked retrieval. *Cognition*, *193*. doi: 10.1016/j.cognition.2019.104027

Stasenko, A., Jacobs, D.M., Salmon, D.P., & Gollan, T.H. (2019). The multilingual naming test (MINT) as a measure of picture naming ability in Alzheimer's disease. *Journal of the International Neuropsychological Society : JINS*, *25*, 821–833. doi: 10.1017/S1355617719000560

Stiver, J., Staffaroni, A.M., Walters, S.M., You, M.Y., Casaletto, K.B., Erlhoff, S.J., … Kramer, J.H. (2021). The Rapid naming test: development and initial validation in typically aging adults. *The Clinical Neuropsychologist*, *0*, 1–22. doi: 10.1080/13854046.2021.1900399

Talamas, A., Kroll, J.F., & Dufour, R. (1999). From form to meaning: stages in the acquisition of second-language vocabulary. *Bilingualism: Language and Cognition*, *2*, 45–58. doi: 10.1017/S1366728999000140

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*, 516–536. doi: 10.1017/S1366728918000421

Umbel, V.M., Pearson, B.Z., Fernández, M.C., & Oller, D.K. (1992). Measuring Bilingual children's receptive vocabularies. *Child Development*, *63*, 1012–1020. doi: 10.1111/j.1467-8624.1992.tb01678.x

Van Assche, E., Duyck, W., & Gollan, T.H. (2013). Whole-language and item-specific control in bilingual language production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1781–1792. doi: 10.1037/a0032859

Wood, C., Hoge, R., Schatschneider, C., & Castilla-Earls, A. (2018). Predictors of item accuracy on the Test de Vocabulario en Imagenes Peabody for Spanish-English speaking children in the United States. *International Journal of Bilingual Education and Bilingualism*, *0*, 1–15. doi: 10.1080/13670050.2018.1547266

Woodcock, R., Muñoz-Sandoval, A., Ruef, A., & Alvarado, C. (2005). *Woodcock-Muñoz language survey-revised*. Itasca, IL: Riverside Publishing.

Zirnstein, M., van Hell, J.G., & Kroll, J.F. (2018). Cognitive Control ability mediates prediction costs in Monolinguals and Bilinguals. *Cognition*, *176*, 87–106. doi: 10.1016/j.cognition.2018.03.001

APPENDIX A. ORAL PROFICIENCY INTERVIEW QUESTIONS AND SCORING RUBRIC

Bilinguals were interviewed with one set in each language counterbalanced between participants with respect to assignment to language and testing order

English Set 1:

A) Where did you grow up? How is it different from or similar to San Diego?

B) **[COOKIE-THEFT PICTURE]:** Please take a look at this picture. Examine the whole picture and then describe everything that you see happening. Be sure to mention all the different things.

C) Tell me about your childhood. What was it like? What do you remember most about it?

D) Tell me about your schedule for the rest of the day. Where will you be and what will you be doing at each hour?

E) There is a debate on whether to extend the school day for children in the USA for the purpose of improving academic performance nationwide. Do you think this is a good or a bad idea and why? How would you defend the opposing view as well?

F) Recently, the state of California passed a law that requires all school-aged children to be vaccinated or have a medical exemption in order to be enrolled in school. Do you think it's the government's place to tell parents whether they should have their children vaccinated? Please explain your reasoning. How would you defend the opposing view as well?

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

English Set 2:

G) Where are you from? How did you learn the languages that you speak?

H) **[BROKEN WINDOW PICTURE]:** Please take a look at this picture. Examine the whole picture and then describe everything that you see happening. Be sure to mention all the different things.

I) Tell me about your time as a student in school. What do you remember most about that experience?

J) Tell me about what you will do next week. Where will you be and what will you be doing each day?

K) Some parents think that bilingual children will not do as well in school as monolingual children. Others say bilingualism is an advantage. What do you think? How would you try to convince someone that your view is the right one?

L) How important are free speech and freedom of the press to a healthy society? Please share your opinion. How would you defend the opposing view as well?

Spanish Set 1:

A) ¿En dónde te criaste? ¿Y cuáles son las diferencias y semejanzas de ese lugar con San Diego?

B) **[COOKIE-THEFT PICTURE]:** Por favor mira esta imagen. Examina la imagen por completo y después describe todo lo que ves que esté sucediendo. Asegúrate de mencionar todas y cada una de las cosas que ves.

C) Cuéntame sobre tu niñez. ¿Cómo fue? ¿Y qué es lo que más recuerdas?

D) Cuénteme sobre lo que tienes programado para cada hora durante el resto del día. ¿En dónde vas a estar y que estarás haciendo?

E) Actualmente hay un debate acerca de extender el día escolar para los alumnos en los Estados Unidos para mejorar el rendimiento académico a nivel nacional. ¿Crees que es una buena o mala idea y por qué? ¿Y cómo defenderías el punto de vista opuesto?

F) Recientemente el estado de California pasó una ley que requiere que todos los alumnos estén vacunados o tengan una excepción médica para poder ingresar en la escuela. ¿Crees que el gobierno debe decidir por los padres si sus hijos deberían de estar vacunados? Por favor explica tu razonamiento. ¿Y cómo defenderías el punto de vista opuesto?

Spanish Set 2:

G) ¿De dónde eres? ¿Cómo aprendiste los idiomas que hablas?

H) **[BROKEN WINDOW PICTURE]:** Por favor mira esta imagen. Examina la imagen por completo y después describe todo lo que ves que esté sucediendo. Asegúrate de mencionar todas y cada una de las cosas que ves.

I) Cuéntame sobre cuando eras un estudiante en la escuela primaria. ¿Qué es lo que más recuerdas de esa experiencia?

J) Cuéntame sobre lo que harás la semana que viene. ¿En dónde vas a estar y que estarás haciendo en cada día?

K) Algunos padres piensan que los niños bilingües no prosperan tanto en la escuela que los niños monolingües. Otros dicen que el ser bilingüe es una ventaja. ¿Qué piensas? ¿Y cómo intentarías convencer a alguien de que tu punto de vista es el correcto?

L) ¿Qué tan importante crees que es la libertad de expresión y de prensa para una sociedad saludable? Por favor comparte tu opinión. ¿Y cómo defenderías el punto de vista opuesto?

Speaking Proficiency Rating Scale

1 = Novice Low = No real functional ability. Given lots of time and cues may be able to exchange greetings, give identity and name a number of familiar objects. Cannot participate in a true conversational exchange.

2 = Novice Middle = Can communicate only very minimally and with great difficulty using a number of isolated words and memorized phrases.

3 = Novice High = Can communicate with some success about simple topics only. Heavy reliance on memorized phrases, or on words provided by person speaking with. Speaks in short or incomplete sentences, and frequent miscommunications occur.

4 = Intermediate Low = Can successfully handle a limited number of uncomplicated communicative tasks by combining and recombining into short statements what they know and what the person speaking with says.

5 = Intermediate Middle = Can successfully handle a variety of uncomplicated communicative tasks about simple topics (food, travel, family, daily activities, and personal preferences). Speaks in full sentences and even with some strings of sentences.

6 = Intermediate High = Can successfully handle many uncomplicated tasks and social situations requiring an exchange of basic information related to work, school, recreation, particular interests, and areas of competence. Some hesitation, errors, and gaps in communication may still occur.

7 = Advanced Low = Can participate actively in most informal and a limited number of formal conversations on activities related to school, home, and leisure activities and, to a lesser degree, those related to events of work, current, public, and personal interest or individual relevance. Can rarely function at the level of formal or professional language and cannot speak at a professional level for an extended period of time.

8 = Advanced Middle = Can handle with ease and confidence a large number of communicative tasks such as informal and some formal exchanges on a variety of concrete topics relating to work, school, home, and leisure activities, as well as to events of current, public, and personal interest or individual relevance. Can sometimes function at a formal or professional level of language but not consistently and not with a broad range of topics.

9 = Advanced High = Can participate fully and effectively in conversations on a variety of topics in formal and informal settings from both concrete and abstract perspectives. Can speak at a formal or professional level of language usually without difficulty. When speaking at a formal or professional level, some patterns of errors may still appear, but these do not interfere with communication.

10 = Superior = Speaks like a highly educated native speaker. Can participate fully and effectively in conversations on a variety of topics in formal and informal settings from both concrete and abstract perspectives with accuracy and fluency using formal and professional quality language. Occasional errors may still occur, but these do not interfere with communication.

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

APPENDIX B

Table B1. List of MINT Sprint Items and Accuracy in Each Language

	MINT Sprint Item & Alternative correct responses		Proportion Correct	
Item Number	English	Spanish	English	Spanish
1	dog	perro	1.00	1.00
2	hand	mano	.99	1.00
3	door	puerta	1.00	.99
4	horse	caballo	1.00	1.00
5	apple	manzana, manzanita	1.00	1.00
6	book	libro	1.00	.99
7	fish	pez, pescado	.99	.99
8	sun	sol	1.00	1.00
9	key	llave	1.00	.99
10	bed	cama	1.00	1.00
11	tree	árbol	1.00	.99
12	chair	silla	1.00	.98
13	moon	luna	.98	.99
14	watch	reloj	.95	.98
15	cake	pastel, tarta, pastelito, torta, bizcocho, ponqué, panqué	1.00	1.00
16	grapes	uvas	1.00	.99
17	scissors	tijeras	1.00	.99
18	airplane, jet, plane, aeroplane	avión, aeroplano, avioneta	1.00	.95
19	fork	tenedor	1.00	.99
20	witch	bruja	1.00	.99
21	glove	guante	.93	.93
22	bear	oso	1.00	1.00
23	hat	sombrero	1.00	.83
24	bone	hueso	1.00	.96
25	iron	plancha	.89	.88
26	basket	canasta, cesta	.98	.81
27	candle	vela, veladora, candela	.98	.83
28	grater, cheese grater	ralladora, rallador, ralladora de queso	.69	.29
29	king	rey	1.00	1.00
30	butterfly	mariposa	1.00	.96
31	tie, necktie	corbata	1.00	.76
32	cloud	nube	1.00	.91
33	leaf, leaves	hoja	1.00	.78
34	clown, joker	payaso, payasito, guasón	.99	.98
35	nurse	enfermera	1.00	.86
36	zipper	cierre, cremallera, zíper	.96	.66
37	bridge	puente	.99	.78
38	feather	pluma	.95	.79
39	drum	tambor, bateria, tambora, tamborín	.96	.83
40	lightbulb, electric bulb, bulb	foco, bombilla/o, ampolleta	.81	.53
41	pacifier, binky	chupón, chupete	.61	.63
42	rainbow	arco iris	1.00	.86
43	nest, bird nest	nido	.98	.50
44	cage, bird cage	jaula	.96	.56
45	lock, padlock, combination lock, combo lock	candado	.99	.51
46	crib, cradle	cuna	.85	.64
47	arrow	flecha	.96	.55
48	radish, beet	rábano, betabel, betarraga, remolacha	.83	.50
49	whale	ballena	.99	.79
50	screw	tornillo, chilillo	.63	.61
51	pomegranate	granada	.78	.58
52	scarf	bufanda, chalina	.98	.68

(*Continued*)

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

Table B1. (*Continued*)

	MINT Sprint Item & Alternative correct responses		Proportion Correct	
Item Number	English	Spanish	English	Spanish
53	saw	serrucho, sierra	.76	.28
54	wig	peluca	.90	.76
55	flippers, fins	aletas	.59	.13
56	kite	papalote, cometa, barrilete, volantín, chichigua, chiringa, piscucha	.94	.41
57	thimble	dedal	.21	.05
58	parachute	paracaídas	.86	.33
59	well, wishing well, water well	pozo, pozo de agua, pozito, aljibe	.81	.29
60	plug	enchufe	.54	.44
61	snail	caracol, baboso	.98	.54
62	crossbow	ballesta	.34	.01
63	dustpan	recogedor, recogedor de basura, pala de residuos	.44	.64
64	flashlight	linterna, lámpara portátil, lámpara de mano	.89	.23
65	peacock	pavo real	.78	.31
66	blind	persiana	.58	.21
67	pitcher, jug	cantaro, jarra, jarro, jarrón, jarrito	.51	.63
68	rake	rastrillo	.75	.15
69	seesaw, teeter-totter	subibaja, balancín, sube y baja, bimbalete, cachumbambé	.73	.35
70	funnel	embudo	.71	.06
71	hinge, door hinge	bisagra	.55	.06
72	gauge, barometer, manometer	medidor, calibrador, barómetro, manometro	.19	.08
73	axle	eje	.25	.00
74	periscope	periscopio	.00	.00
75	mortar or pestle	mortero, molcajete o mano de mortero	.09	.49
76	metronome	metrónomo	.16	.01
77	anvil	yunque	.26	.03
78	gyroscope	giroscopio	.04	.00
79	bellows	fuelle	.00	.00
80	porthole	portilla, ojo de buey, escotilla	.06	.03

APPENDIX C

Table C1. Correlations between English proficiency measures

	Oral Proficiency Interview	Original MINT items	MINT Sprint 2nd Pass	WMPV	WMSC	Average Self-Rating	Spoken Self-Rating[b]
Original MINT	.431**	1					
MINT Sprint 2nd Pass	.455**	.980**	1				
WMPV	.469**	.774**	.809**	1			
WMSC	.235*	.380**	.382**	.383**	1		
Average Self-Rating	.447**	.490**	.472**	.446**	.251*	1	
Spoken Self-Rating	.462**	.501**	.481**	.476**	.301**	.940**	1
Translation Recognition Accuracy[a]	.017	.152	.137	.161	.277*	.080	.055

** $p < .01$, * $p < .05$. WMPV, Woodcock-Muñoz Picture Vocabulary; WMSC, Woodcock-Muñoz Sentence Completion.

[a] The translation recognition task only included an overall accuracy score (i.e., no separate scores for each language).

[b] $n = 79$, one bilingual left the self-rating for spoken proficiency in Spanish blank

Table C2. Correlations between Spanish proficiency measures

	Oral Proficiency Interview	Original MINT Items	MINT Sprint 2nd Pass	WMPV	WMSC	Average Self-Rating	Spoken Self-Rating[b]
Original MINT	.659**	1					
MINT Sprint 2nd Pass	.662**	.990**	1				
WMPV	.650**	.757**	.743**	1			
WMSC	.576**	.729**	.712**	.628**	1		
Average Self-Rating	.421**	.394**	.399**	.395**	.294**	1	
Spoken Self-Rating	.394**	.375**	.385**	.400**	.205	.884**	1
Translation Recognition Accuracy[a]	.318**	.425**	.408**	.489**	.382**	.153	.169

** $p < .01$, * $p < .05$. WMPV, Woodcock-Muñoz Picture Vocabulary; WMSC, Woodcock-Muñoz Sentence Completion.

[a] The translation recognition task only included an overall accuracy score (i.e., no separate scores for each language).

[b] $n = 79$, one bilingual left the self-rating for spoken proficiency in Spanish blank

Table C3. Correlations between language dominance scores across different proficiency measures

	Oral Proficiency Interview	Original MINT	MINT Sprint 2nd Pass	WMPV	WMPC	Average Self-rating	Spoken Self-Rating[b]
Original MINT	.773**	1					
MINT Sprint 2nd Pass	.780**	.986**	1				
WMPV	.740**	.809**	.821**	1			
WMPC	.435**	−.538**	.534**	.530**	1		
Average Self-rating	.569**	.630**	.643**	.614**	.452**	1	
Spoken Self-Rating	.537**	.590**	.594**	.569**	.345**	.865**	1
Translation Recognition Accuracy[a]	−.293**	−.282**	−.237*	−.213	−.129	−.087	−.128

** $p < .01$, * $p < .05$. WMPV, Woodcock-Muñoz Picture Vocabulary; WMPC, Woodcock-Muñoz Passage Comprehension.

[a] The translation recognition task only included an overall accuracy score (i.e., no separate scores for each language).

[b] $n = 79$, one bilingual left the self-rating for spoken proficiency in Spanish blank

Table C4. Correlations between bilingual index scores across different proficiency measures

	Oral Proficiency Interview	Original MINT	MINT Sprint 2nd Pass	WMPV	WMPC	Average Self-Rating	Spoken Self-Rating[b]
Original MINT	.694**	1					
MINT Sprint 2nd Pass	.691**	.974**	1				
WMPV	.455**	.542**	.582**	1			
WMPC	.287**	.324**	.339**	.351**	1		
Average Self-Rating	.393**	.332**	.331**	.316**	.186	1	
Spoken Self-Rating	.298**	.209	.209	.242*	.191	.800**	1
Translation Recognition Accuracy[a]	.323**	.349**	.300**	.098	.124	.181	.195

** $p < .01$, * $p < .05$. WMPV, Woodcock-Muñoz Picture Vocabulary; WMPC, Woodcock-Muñoz Passage Comprehension.

[a] The translation recognition task only included an overall accuracy score (i.e., no separate scores for each language).

[b] $n = 79$, one bilingual left the self-rating for spoken proficiency in Spanish blank

https://doi.org/10.1017/S1355617721001004 Published online by Cambridge University Press

APPENDIX D. TRANSLATION RECOGNITION TASK RESULTS

Table D1. Mean response times in (in milliseconds) and error rate proportion in the translation recognition task[a]

		Reaction Time		Proportion of Errors	
Decision Type	Condition	*M*	*SD*	*M*	*SD*
Yes decisions	Translation Pairs	711	149	.05	.03
No Decisions	Form Pairs				
	Related	886	207	.07	.07
	Unrelated	827	182	.01	.02
	difference	*59*	*121*	*.06*	*.08*
	Semantic Pairs				
	Related	1040	321	.25	.12
	Unrelated	836	214	.02	.04
	difference	*205*	*212*	*.23*	*.11*

[a] A 2x2 repeated measures ANOVA with RTs on No decisions as the dependent variable and distractor type (form or semantic) and relatedness (related or unrelated) as independent variables. Bilinguals took longer to reject semantic than form distractors, a main effect of distractor type ($F(1,79) = 30.4$, $p < .001$, $\eta_p^2 = .278$, $MSE = .531$), and longer to reject related than unrelated distractors, a main effect of relatedness ($F(1,79) = 92.3$, $p < .001$, $\eta_p^2 = .539$, $MSE = 1.39$), and semantic distractors slowed responses much more than form distractors (characteristic of proficient bilinguals; Talamas et al., 1999), an interaction between distractor type and relatedness ($F(1,79) = 28.7$, $p < .001$, $\eta_p^2 = .266$, $MSE = .424$).