Running-head: When A Seven is Not a Seven

When a Seven is Not a Seven: Self-Ratings of Bilingual Language Proficiency Differ Between and Within Language Populations*

Brendan Tomoschuk
University of California, San Diego

Victor S. Ferreira
University of California, San Diego

Tamar H. Gollan
University of California, San Diego

* Acknowledgements: The authors thank Hal Pashler and Mark Appelbaum for helpful discussion, and Rosa Montoya, Mayra Murillo and Tiffany Ho for gathering and coding the data. Funding: This research was supported by grants from the National Institute on Deafness and Other Communication Disorders (011492), the National Institute of Child Health and Human Development (050287, 051030, 079426) and the National Science Foundation (BCS1457159).

Address for Correspondence:
Brendan Tomoschuk, Department of Psychology, University of California, San Diego,
9500 Gilman Drive, La Jolla, CA, 92093-0109
E-mail: btomoschuk@ucsd.edu

Keywords: bilingualism, language dominance, self-ratings, MINT, Oral Proficiency Interview

Abstract

Self-ratings of language proficiency are ubiquitous in research on bilingualism, but little is known about their validity, especially when the same scale is used across different types of bilinguals. Self-ratings and picture naming data from 1044 Spanish-English and 519 Chinese-English bilinguals were analyzed in five between- and within-population comparisons. Chinese-English bilinguals scored more extremely than Spanish-English bilinguals, and in opposite directions at different endpoints of the self-ratings scale. Regrouping bilinguals by dominant language, instead of language membership, reduced discrepancies but significant group differences remained. Population differences appeared even in English, though this language is shared between populations. These results demonstrate significant problems with self-ratings especially when comparing bilinguals of different language combinations, and subgroups of bilinguals who speak the same languages but vary in acquisition history and/or dominance. Objective proficiency measures (e.g., picture naming or proficiency interviews) are superior to self-ratings, to maximize classification accuracy and consistency across studies.

When a Seven is Not a Seven:

Self-Ratings of Bilingual Language Proficiency Differ Between and Within Language Populations

Language proficiency is a uniquely important variable in bilingual research. It affects how quickly and effectively bilinguals can access words in their languages, how easily they control language choice and output, and many other phenomena that have implications for understanding linguistic behavior more generally. It is therefore important for researchers to measure a bilingual's language proficiency in the most accurate way possible.

Proficiency is most often measured by self-ratings (Li, Sepanski & Zhao, 2006). Participants are asked to report how well they read, write, speak or comprehend spoken language, typically on a scale of 1 to 7 (or 1 to 10) with 1 representing not at all proficient in a language and 7 being a native speaker of that language. These self-ratings are simple to collect and record. Unfortunately, this simplicity comes with some drawbacks. Self-ratings are vulnerable to the subjectivity and variability of the participants who provide them as well as the way researchers frame the questions and the experiment more broadly (see Schwarz, 1999; Dunn & Foxtree, 2009). Zell and Krizan (2014), in particular, examined the relationship between self-evaluations and performance measures across 22 meta-analyses and found that there was only a moderate correlation between the two ($M = 0.29$, $SD = 0.11$).

A related and ongoing discussion within the field has been lack of consistency across researchers in how self-ratings are collected (see Grosjean, 1998). For this reason, some investigators have developed standardized language history questionnaires, with the intent of reducing between-study variability. One of the most commonly used was developed by Marian, Blumenfeld and Kaushanskya (2007), who standardized self-rated proficiency questions and

explored the relationship between language background and objective measures of bilingual language proficiency. They administered their questionnaire and a battery of objective proficiency measures (picture naming, passage comprehension, reading fluency, sound awareness and grammaticality judgment) in two different multilingual populations and used a principal components analysis to identify several factors of note when using language background to predict proficiency. In a factor they called "relative L2-L1 competence," they found that self-rated proficiency of the non-dominant language and estimated current language use combine to account for the most variance (about 25%) in predicting objective proficiency. Many bilingual studies use these results to justify the use of self-ratings, but do not also consider estimated daily language use, acquisition history or other factors the LEAP questionnaire examined.

Although self-ratings are simple and a standardized questionnaire can increase consistency between labs and across experiments relative to not measuring proficiency at all, speakers of different languages can be very different in terms of their linguistic profiles. Languages differ from one another in structure and form, and the people that speak them come from different cultures in which similarly worded questions can take on different meanings. Even within a bilingual language population, some bilinguals may have learned and constantly use both languages at home and at work, while others might have learned one language first and use different languages at home or in school or work, causing language proficiency to vary by setting. Grosjean (1998) describes this difference as the *complementarity principle*, stating that "bilinguals are rarely equally fluent in all language skills in all languages." These and other cultural and personal differences can affect language proficiency and dominance, which could in turn affect how proficiency is self-rated. It seems unlikely, therefore, that bilinguals from diverse

backgrounds would factor all of this variation into a one-dimensional rating of their abilities in each language in the same way. Despite these drawbacks, many researchers still opt for self-ratings rather than objective proficiency. Hulstijn (2012) reports that 55% of 140 empirical studies published in *Bilingualism: Language and Cognition*, did not measure language proficiency objectively.

In this paper we hope to demonstrate the importance of factoring objective measures of proficiency into studies of bilingualism. One such objective measure is the Multilingual Naming test or MINT. The MINT is a standardized picture-naming task in which participants name 68 pictures of varying frequency in both of their languages. It has been validated as a proficiency measure that captures variance in lexical retrieval for bilinguals who speak English, Spanish, and Mandarin (Gollan, Weissberger, Runnqvist, Montoya & Cera, 2012; Ivanova, Salmon & Gollan, 2012; Sheng, Lu & Gollan, 2014), and also appears to function similarly for predicting proficiency in Hebrew-English, Spanish-English and Chinese-English bilingual children and young adults (Gollan, Starr, & Ferreira, 2015). The MINT excludes cognates (translations that are phonologically similar between the two target languages), and words with potential cultural differences (such as *abacus* which is low frequency in English but higher in Mandarin since it is used as an educational tool in China). While not a catch-all measure of all domains that affect language proficiency (including grammar and syntax), it was developed and measured against the more comprehensive Oral Proficiency Interviews (OPI), and demonstrated to be more accurate than the Boston Naming Task (BNT, Kaplan, Goodglass & Weintraub, 1983) for capturing bilingual language proficiency. Here we seek to further improve consistency across studies in bilingual research by investigating how effective subjective metrics like self-rated

proficiency are at capturing similarities and differences between language combinations, and how well these relate to the MINT.

In the present study, we performed five analyses on two pooled sets of data from previous studies that used the MINT, to measure the extent to which self-rated proficiency scores can reasonably be compared or collapsed across Spanish-English (typically people who grew up in the greater San Diego area) and Chinese-English (people who grew up in China studying at UC San Diego, or Chinese heritage speakers who grew up in the U.S.) bilinguals and with different dominance profiles (English-dominant or other-language dominant, see Table 1 for full participant information). For each of these analyses we investigated this relationship in self-reports of English as well as a bilingual's other language. We also report a simulation that explores the effects suggested by these analyses. One hypothesis is that the simple nature of self-rated proficiency is enough to allow bilinguals to reasonably estimate their own skills and that this estimation will allow for valid comparison between bilingual populations and within-language subgroups. If so, we should see that the relationship between the self-ratings and MINT scores pattern together regardless of bilingual population (Analyses 1 and 2) and within-language subgroups (Analyses 3, 4 and 5). Alternatively, different bilingual sub-groups may rate themselves based on distinct subjective standards, for example, assessing their own performance against different comparison groups. If so, between-group comparisons could reveal substantial differences across groups in chosen self-rating level and objectively measured performance. The latter pattern would raise significant concerns with the use of self-ratings to measure proficiency when comparing or collapsing across bilinguals of different language combinations or even dominance profiles within bilinguals of just one group.

Analysis 1: Self-ratings and Language combination

To examine consistency in self-rated language proficiency between populations, we first looked at MINT scores as a function of self-rated proficiency in both languages, split into Spanish-English bilinguals and Chinese-English bilinguals.

Method

Participants. Spanish-English (n = 992) and Chinese-English (n = 223) bilingual undergraduates at the University of California, San Diego participated in 15 different studies for course credit. All Spanish-English bilinguals reported proficiency in Spanish and English with 702 reporting English as their dominant language, 128 reporting Spanish, and 162 reporting balanced proficiency. All Chinese-English bilinguals reported proficiency in both Mandarin and English with 72 reporting English as their dominant language, 139 reporting Mandarin as their dominant language, and 12 reporting balanced proficiency. Full participant characteristics are listed in Table 1.

<Insert Tables 1a and 1b about here>

Procedure. Bilinguals completed a language history questionnaire in which they rated their proficiency in both languages (and any other they reported knowing) on speaking, reading, writing, and listening on a scale from 1 to 7, with the following anchors: 1 - Almost none, 2 – Very Poor, 3 – Fair, 4 – Functional, 5 – Good, 6 – Very Good, 7 – Like a native speaker. In most cases, bilinguals completed the questionnaire at the beginning of the experiments and the MINT (Gollan et al., 2012) at the end, first in English and then in Spanish or Mandarin. Forty of the Spanish-English bilinguals completed their language history questionnaire at the end of the experiment, after the MINT.

Analysis. Simple regression was done using the Stats package in R (R Core Team, 2013). Self-rated speaking proficiency was the independent variable and MINT scores first with either Mandarin or Spanish, and then again with English, were the dependent measures. In this way, self-rated speaking accounts for as much of the variance as possible before the factors of interest are considered.

Results and discussion

Figure 1 illustrates the results of these first analyses with Figure 1a showing the other-language results, and Figure 1b showing English. Figure 1a reveals a crossover interaction showing that, on average, Chinese-English bilinguals obtained higher other-language MINT scores at higher self-ratings and lower MINT scores at lower ratings as compared to Spanish-English bilinguals. To illustrate, Chinese-English bilinguals who rated themselves a 7 (out of 7) in Chinese proficiency scored an average of 59.0 (6.1) out of 68 on the Chinese MINT whereas Spanish-English bilinguals who rated themselves as a 7 in Spanish proficiency scored 50.9 (8.0) out of 68, that is, greater than a standard deviation difference across language combinations. Conversely, for the bilinguals who rated themselves a 3, Chinese-English bilinguals averaged 30.1 (12.0) out of 68 while Spanish-English bilinguals averaged 42.1 (9.9) out of 68, an even larger difference. Though there are considerably fewer data points at the low than at the high end of the self-rating scale, particularly for Chinese-English bilinguals, these differences resulted in a significant interaction between self-rated proficiency and language combination, as shown in Table 2.

Furthermore, Figure 1b also shows a significant interaction (analyses reported in Table 3) between English self-rated speaking and English MINT scores such that Spanish-English bilinguals scored higher in the MINT at any given self-rating as compared to Chinese-English

bilinguals, except at the higher end of the scale. This may suggest that Spanish-English bilinguals had higher standards of performance in both languages, but this cannot account for the cross-over pattern found in Figure 1a. Population differences in self-rating, especially in the language both bilingual populations share (English, in this case), could introduce potentially serious problems in studies that use self-ratings to select proficient bilinguals.

Why might self-rating differences arise between bilinguals of different language combinations? It may be that Chinese-English bilinguals perform more extremely at either end of the self-rated proficiency scale (when rating Chinese), simply due to linguistic differences between the Chinese and Spanish languages, or cultural differences between the populations. Alternatively, it may be that other common factors of bilingualism research (such as first versus second language dominant bilinguals, age of acquisition) may drive this population level effect. Before considering these options, it is important to confirm that the MINT converges across languages with other objective measures of proficiency to a greater extent than with the self-ratings.

<Insert Figure 1 about here>

<Insert Tables 2 and 3 about here>

Analysis 2: MINT Validation

One reason why scores might differ across populations is if the MINT itself introduces a between-population bias. To assess this empirically, in Analysis 2 we examined the validity of the MINT by reanalyzing data from Gollan et al. (2012) and Sheng et al. (2014) together to provide direct comparison of self-rated proficiency across the two different language combinations (something that was not done in the original MINT papers). These studies investigated the validity of the MINT, in English and either Spanish or Chinese by comparing

MINT scores to Oral Proficiency Interview (OPI) scores. OPI scores are proficiency ratings given by a single experimenter who is trained to look for specific criteria when determining proficiency level based on a structured face-to-face interview in each language. These interviews were modeled on methods developed by the American Council for Teaching Foreign Languages (ACTFL; see Gollan et al., 2012). Participant characteristics are listed in Table 4.

Method

Participants. Data from 52 Spanish-English bilinguals and 62 Chinese-English bilinguals were reanalyzed from Gollan et al. (2012) and Sheng et al. (2014), respectively.

<Insert Table 4 about here>

Procedure. The procedures were identical to those described in Analysis 1.

Analysis. The analysis differed only in that OPI scores were used instead of self-rated speaking proficiency as a predictor of MINT scores. Additionally, in these data, the MINT score is reported as a proportion correct (with a score of 1 meaning that all 68 pictures of the MINT were named correctly). This was done because the original MINT data were compared with the Boston Naming Test (Kaplan et al. 1983), which has a different total number of pictures than the MINT.

Results and discussion

Figure 2a shows Spanish/Chinese MINT scores and Figure 2b shows English MINT scores from Gollan et al. (2012) and Sheng et al. (2014) as predicted by the Oral Proficiency Interview. The English OPI and the English MINT were positively correlated (0.47, $t = 5.69$, $p < 0.001$) as were the other language OPI and the other language MINT (0.72 ,$t = 10.93$, $p < 0.001$), showing that both are closely related regardless of language combination. Although the correlations between OPI and MINT range from moderate to high, these data showed no

interaction between the OPI and language combination in either language (model results detailed in Tables 5 and 6). Thus, this analysis supports the internal validity of the MINT, and further suggests that the real source of discrepancy between bilinguals of different language combinations in Analysis 1 were biases in the self-ratings.

As such, it seems that the MINT successfully does what it was designed to do – that is, it is equally successful in measuring proficiency across bilinguals of different language-combinations and does not vary significantly as a measure between these two language populations. Note that although the OPI involves subjective rating (as do self-ratings), ratings are made by a single trained interviewer with consistent criteria for all bilinguals participating in the study – whereas each self-rating is assigned by a different individual (the subject him or herself) who might have different standards of performance and reference frame for determining proficiency level.

<Insert Figure 2 about here>

<Insert Tables 5 and 6 about here>

Analysis 3: Exploring Language Dominance

Analysis 1 showed a crossover interaction such that Chinese-English bilinguals rated themselves more extremely in Chinese at both ends of the proficiency scale relative to Spanish-English bilinguals in Spanish. It likewise showed that Spanish-English bilinguals score higher in the English MINT than Chinese-English bilinguals at any given self-rating, except for the very highest rating where the two groups converged. To explore what factors within a population might drive these differences, we calculated and use two measures of dominance to understand how dominance might have affected the results of Analysis 1. For Analysis 3.1, we split each language population into three groups based on self-reported language dominance including

English-dominant, other-language dominant, and balanced bilinguals (i.e., those who rate their skills in both languages as the same on average across four modalities). In Analysis 3.2 we calculated dominance on a continuous scale using the Edinburgh handedness method, as explained below, and described in Chapter 5 (Birdsong) of Silvá-Corvalan and Treffers-Daller (2016).

Analysis 3.1

In this part of the analysis, we use participants' own self-ratings to determine language dominance, in accordance with self-rating. Self-rated balanced bilinguals were included in Analyses 1-2, but were omitted from Analysis 3.1 – only 12 Chinese-English bilinguals rated themselves as balanced, and so their omission is unlikely to substantively influence the results. Though not included in the model, data for self-rated balanced Spanish-English bilinguals are included in Figure 3 to illustrate how they differed from the other groups. Note that despite their balanced ratings, 59.2% of these bilinguals named at least 10% more words in one language or the other in the MINT, making them unbalanced bilinguals by this objective measure (see General Discussion). In Gollan et al. (2012) and Sheng et al. (2014), a more conservative 5% margin was used in considering bilinguals to be balanced, and with this margin 77.8% of Spanish-English bilinguals in this study who rated themselves as balanced nevertheless produced more pictures in one than the other language (88% of this subset obtaining higher scores in English than Spanish).

Method

The participants, procedure and data were all identical to Analysis 1, except for the exclusion of balanced bilinguals (as noted above). Multiple regression was used in this analysis,

in contrast to the simple regression in Analysis 1. Bilingual population and language dominance were used as predictor variables.

Self-assessed language dominance was determined by averaging all four ratings for each modality in each language and taking whichever language had the higher average self-rated proficiency to be the dominant language. If these averages were equal, the bilingual was considered to be self-rated as balanced. Because the MINT measures productive vocabulary, this analysis was also redone using only self-rated speaking to determine self-assessed language dominance (for discussion on assessment of dominance see Silva-Corválan & Treffers-Daller, 2016). Statistical differences between these two methods are noted in the results.

In pairwise comparisons, all Spanish-English dominance groups as determined by self-ratings differed from one another in MINT scores. That is to say, English-dominant bilinguals had significantly higher English MINT scores than balanced bilinguals, who had significantly higher English MINT scores than Spanish-dominant bilinguals. Similarly, Spanish-dominant bilinguals had significantly higher Spanish MINT scores than balanced bilinguals, who had significantly higher Spanish MINT scores than English-dominant bilinguals. Likewise English-dominant Chinese-English bilinguals had significantly higher English MINT scores than Chinese-dominant bilinguals. Chinese-dominant bilinguals, in turn, had significantly higher Chinese MINT scores than English-dominant Chinese-English bilinguals. All effects were significant and robust. As in the main finding of this analysis, balanced Chinese-English bilinguals were omitted due to a low n.

Results and discussion

Figure 3a shows the results in the other-language group, while Figure 3b shows the results in English. Analysis of other-language performance revealed a significant interaction

between language dominance and self-rated proficiency, regardless of language combination (model results in Table 7), such that dominance alone drove a difference in the relationship between the MINT scores and self-rated proficiency; bilinguals in the English-dominant group performed worse in their respective other-language MINT at any given self-rating than other-language dominant bilinguals. The reverse was true for English MINT scores and self-ratings. English-dominant groups scored higher in English at any given self-rating as compared to their other-language dominant counterparts (model results in Table 8).[1]

There was also a significant three-way interaction in the English MINT (Figure 3b, Table 8). Chinese-dominant bilinguals scored worse than Spanish-dominant bilinguals in English while the opposite was true for their English-dominant counterparts. That is, English-dominant Chinese-English bilinguals scored better on average than English-dominant Spanish-English bilinguals in English. Pairwise comparisons showed that the differences between English-dominant subgroups was significant ($F(1,770) = 5.20$, $p = .023$), and the difference between other-language dominance subgroups was not ($F(1,263) = 0.14$, $p = .72$).

These data suggest that the crossover interaction seen in Figure 1a is in part driven by dominance groups within bilingual populations seen in Figure 3a, demonstrating that some of the population level differences can be explained by the within-language group factors not usually considered in bilingual research. More specifically (and assuming differences in power are not responsible for the significance of one but not the other pairwise comparisons), across bilinguals of different language combinations, bilinguals not dominant in English seem to assign self-ratings based on more similar points of comparison across different language combinations. However, among English-dominant bilinguals, the Chinese speakers may have overestimated their abilities in Chinese and underestimated their abilities in English, or the Spanish speakers

may have underestimated their abilities in Spanish and overestimated their abilities in English, or both.

<Insert Figure 3.1 about here>

<Insert Tables 7 and 8 about here>

Analysis 3.2

In this analysis, the Edinburgh method was used to calculate dominance. The Edinburgh method is calculated:

$$\frac{\text{Language A MINT} - \text{Language B MINT}}{\text{Language A MINT} + \text{Language B MINT}}$$

In this case, we consider Language A to be English and Language B to be Spanish or Chinese. This calculation therefore gives a score that is positive to reflect English dominance or a negative to represent other-language dominance. For example, a Spanish-English bilingual who scored a 55 on the English MINT and a 45 on the Spanish MINT would have a dominance score of 0.10 (calculated as (55-45)/(55+45)), while a bilingual who scored a 45 on the English MINT and a 55 on the Spanish MINT would have a dominance score of -0.10. By using this metric, all bilinguals (including those previously categorized as balanced) are able to be factored into the analysis.

Method

The participants, procedure and data are all identical to Analysis 3.1, with the exception that dominance was calculated as a continuous, rather than categorical variable as described. Consequently, bilinguals excluded from analysis 3.1 for being balanced in their self-ratings were included in this analysis.

Results and Discussion

In Analysis 3.1, language dominance was operationalized as a categorical variable determined by self-rating score. In this analysis, the Edinburgh method was used to turn the MINT score data into a continuous measure of dominance. Rather than simply calling a bilingual English or Other-langauge dominant, they were assigned a dominance score where positive numbers indicated English dominance and negative numbers, other-language dominance. Numbers of a greater magnitude show stronger language-dominance. Table 9 shows the regression outcome for predicting other-language MINT scores when using the Edinburgh dominance measure. Every factor yields a significant contribution to the total variance accounted for with the exception of the interaction between self-rated speaking and language dominance. The significance of the three-way interaction between self-rated speaking, language combination and language dominance suggest that these groups do rate themselves differently in their other-language (non-English language) based on their language dominance. Table 10 likewise shows that every factor in the prediction of English MINT scores is a significant contributor to the overall variance accounted for in the model.

There were no major differences in the significance outcomes between categorical (Analysis 3.1) or continuous (Analysis 3.2) measures of dominance. While continuous measures acted as overall better predictors of MINT score, both models suggested that different language combinations perform differently on both the other-language MINT and the English MINT as a fuction of self-rated speaking and language dominance.

<Insert Tables 9 and 10 about here>

Analysis 4: Chinese-English Bilinguals with Different Language Learning History

One possible explanation for the results found in Analysis 3 is that bilinguals do not rate themselves in comparison to every other speaker of that language. For example, Chinese-English

bilinguals raised in the USA may not rate themselves in comparison to Chinese learners or Chinese monolinguals (to name only two similar populations). To explore this possibility, we collapsed the Chinese-English bilinguals analyzed in the previous analyses into one group (referred to as the Chinese-English group). This group was analyzed alongside two other experimental groups run under similar conditions, but recruited for different characteristics: a sample of undergraduates who grew up in the United States but were exposed to Chinese by at least one parent growing up (referred to as Chinese exposed; from Tao, Taft & Gollan, 2015) and a sample of Chinese native speakers (referred to as Chinese immigrated; unpublished data used with permission of Rachel Ostrand) who immigrated to the USA relatively recently (age of arrival: M = 15.5, SD 5.4).

Method

Participants. Table 11 shows participant characteristics for the two new groups of Chinese language users: Chinese exposed undergraduates (N = 90) and recently immigrated Chinese undergraduates (N = 144) who participated in 2 different studies and were analyzed together with the 223 Chinese-English bilinguals from Analysis 1. Recruited for different backgrounds, the three populations differed in their English use growing up. When prompted as part of the language history questionnaire "While you were growing up (from birth through high school), please approximate the percentage of time during an average day that you used each language" the Chinese exposed undergraduates reported an average of 72.3% (17.3) use English, the Chinese-English bilinguals report 33.9% (24.3) use English and the recently immigrated Chinese speakers report 20.1% (18.7) use English. All three of these populations differed significantly from one another in t-tests at p <0.001.

Procedure. The procedure was identical to Analysis 1 with the exception that the recently immigrated Chinese speaker group received an abridged version of the language history questionnaire that only recorded self-ratings for speaking and listening.

Analysis. Data from these groups were analyzed as in Analysis 1. Simple regression was done using self-rated proficiency and Chinese bilingual subgroup as factors in predicting Chinese and English MINT scores.

<Insert Table 9 about here>

Results and discussion

Figure 4 illustrates the results for each of the three groups. The results in Figure 4a reveal a strong between-group difference in self-ratings relative to proficiency on the same MINT tests (model results shown in Table 12). The recently immigrated Chinese speaker group, who had minimal exposure to English, scored the highest on the Chinese MINT at any given self-rating, while the Chinese-English bilinguals from Analyses 1 and 3 scored in the middle, and the Chinese exposed group scored the lowest in the Chinese MINT at any given self-rated proficiency score. In other words, relative to their performance on the Chinese MINT, recently immigrated Chinese speakers tended to provide lower self-ratings, Chinese-exposed speakers tended to provide higher self-ratings, and Chinese-English bilinguals were in the middle. This may be because each population rates themselves relative to their own peers, which would cause recently immigrated speakers to rate themselves lower and Chinese-exposed speakers to rate themselves higher given the same objective level of performance (e.g. recently immigrated speakers are comparing themselves to family and friends in China, while Chinese-exposed speakers are comparing themselves to native English speakers in the US). Similarly, Chinese exposed speakers scored highest in the English MINT, shown in Figure 4b, and only rated

themselves at 6 or 7 in English speaking ability, whereas the other two populations behaved similarly, rating themselves lower in English and also scoring lower in English (model results shown in Table 13).

These data suggest that while every participant was asked the same question ("How well do you rate your Chinese [or English] proficiency"), and took the same MINT tests, the nature of the population and how participants were recruited can impact self-ratings; any given group is likely not considering other, arguably similar groups, or they collapse together groups within their population in their judgment. We might therefore speculate that this difference also accounts for some of the between language-combinations differences, as a Chinese speaker has no internal comparison for how proficient a Spanish speaker might be in Spanish relative to their own proficiency in Chinese.

<Insert Figure 4 about here>

<Insert Tables 12 and 13 about here>

Analysis 5: Languages Grouped by Dominance

Given that the correlation between self-rating and objective measures is typically stronger in the non-dominant language (Marian et al., 2007; Gollan et al., 2012; Sheng et al., 2014) than in the dominant language, here we asked whether the self-ratings are more accurate if we divide them based on dominance rather than by language membership. We therefore collapsed the Spanish-English and Chinese-English populations across languages, and separated their responses into different analyses, one for the self-rated dominant language and another for the self-rated non-dominant language. Thus, Analysis 5 differs from Analysis 3 in that only in Analysis 5 were MINT scores from different tests (English and Spanish or English and Chinese) collapsed together (see below).

Method

The participants, procedure, and data were all identical to Analyses 1 and 3. Responses in this analysis were separated by dominant and non-dominant languages (bilinguals who self-rated themselves as balanced bilinguals were again excluded). Therefore, Chinese MINT scores of self-rated Chinese-dominant bilinguals were grouped for analysis with Spanish MINT scores of self-rated Spanish-dominant bilinguals, and the English MINT scores of self-rated English-dominant Spanish-English bilinguals were grouped with the English scores of self-rated English-dominant Chinese-English bilinguals. Likewise, all non-dominant language responses were grouped together collapsing across language (English, Spanish, or Chinese).

Results and discussion

The results of the dominant language model are plotted in Figure 5a. These show a crossover interaction similar to Analysis 1 such that Chinese-English bilinguals had higher MINT scores than Spanish-English bilinguals at high ends of the self-rating scale, but lower MINT scores on the lower end of the self-ratings scale (see Table 14 for full model results). This was true in their dominant language, regardless of whether or not the dominant language was English, or Spanish/Chinese. Of note, this interaction (Figure 5a) appeared to be numerically smaller than that shown in Figure 1.[2]

Another notable difference was that 28.2% of Spanish-English bilinguals provided a rating of less than 7 for their dominant self-rated speaking proficiency, whereas only 7.6% of Chinese-English bilinguals provided a rating of less than 7. This further demonstrates that these populations behave differently from one another in their methods of self-assessment, and may reflect the fact that a greater proportion of the Spanish-English bilinguals are switched-

dominance bilinguals (they learned and used Spanish-dominantly from birth, but then became English-dominant over time with immersion in an English-dominant environment).

The non-dominant language results are shown in Figure 5b. These show a significant main effect of language combination such that Chinese-English bilinguals scored higher in their non-dominant language than Spanish-English bilinguals, at all points on the self-rating scale (see Table 15 for model results). This is unsurprising given that a greater proportion of Chinese-English bilinguals were not English-dominant, which means they were immersed in their non-dominant language at the time of testing, which would be expected to improve proficiency substantially. Though effects shown in Figure 5 are numerically smaller in size than those shown in Figure 1, the potentially problematic population differences nevertheless remained highly robust and in this case in opposite directions for the dominant versus non-dominant languages.

<Insert Figure 5 about here>

<Insert Tables 14 and 15 about here>

To supplement this analysis (using the same participants from Analyses 1, 3 and 5, detailed in Table 1) and explore bilinguals' ability to self-assess their own language dominance more specifically, we report correlations between self-ratings and self-rated dominance scores (English self-rating minus other-language self-rating) or objectively measured dominance scores (English MINT score minus other-language MINT score). These were done both for self-rating scores including the one used in most analyses above, that is, the average of self-ratings for all four modalities, and the simpler method, using only self-rated speaking scores. These correlations are shown in Table 16.

Though the correlations were statistically robust they appeared to vary considerably between groups. Specifically, it seemed as if Chinese speakers were better at rating their own

proficiency in each language (top rows of Table 16). The variance was higher for Spanish-English responses than Chinese-English responses. This apparent difference between groups disappeared once broken down by dominance (middle rows of Table 16), however, these correlations were relatively weak in both groups. Of interest, and consistent with previous reports (Marian et al., 2007; Gollan et al., 2012; Sheng et al., 2014), dominance scores (dominant minus non-dominant) revealed the highest correlations with objectively measured proficiency. This indicates that bilinguals are much better at rating which of their languages is stronger than they are at rating absolute proficiency level in each language. Finally, there were no striking differences in the size of the correlations when averaging self-ratings from all four modalities versus using just the speaking rating. However, as noted above, hundreds of participants appeared not to have a dominant language when relying only on speaking ratings, thus the average measure might be preferable.

<Insert Table 16 about here>

Other-Language Group Comparison Simulation

To demonstrate how population level differences in self-report judgments might lead to problematic results, we conducted one final analysis. Specifically, we conducted a simulation using participants' other-language self-ratings and MINT scores to explore concerns that might come from relying on self-ratings.

Throughout these analyses, our approach has been to pool many participants from many different studies. An advantage of this approach is that we had hundreds of participants and therefore strong statistical power. A disadvantage is that because bilinguals did many different tasks across different experiments, we don't have any single performance variable (e.g., between-language priming effects) to determine whether relying on self-ratings to make

between-group comparisons can lead to problematic conclusions (relative to relying on an objective measure such as the MINT). And so instead, we conducted a simulation whereby we assigned each participant from Analysis 1 a dummy response time (RT) score that is meant to reflect performance on any task thought to be modulated by (objectively measured) language proficiency. We generated these dummy RTs by random selection from a normal distribution based on their other-language MINT scores. To do this, a participant's MINT score was multiplied by 10. That number was used as the mean of a normal distribution with standard deviation 100, and a value was drawn from that normal distribution. This number was subtracted from 1200 (in order to simulate that higher proficiency leads to faster response times). Finally, this number was assigned as that participant's dummy RT.

For example, a bilingual who scored 60 on the MINT (i.e., who scored well) had a value randomly sampled from a distribution with mean 600 (60 * 10) and standard deviation of 100; this value was subtracted from 1200 and assigned as his or her dummy RT. So, if the randomly selected value for this bilingual were 630, the assigned dummy RT would be 570 ms. Meanwhile, a bilingual who scored 30 on the MINT (i.e., who scored poorly) had a value randomly sampled from a distribution with mean 300 (30 * 10) and standard deviation of 100, with this value subtracted from 1200. If the randomly selected value for this bilingual were 330, the assigned dummy RT would be 870 ms. This will lead on average to slower dummy RTs for bilinguals with lower MINT scores and faster dummy RTs for bilinguals with smaller MINT scores, with a stochastic component (random selection from the normal distribution) to reflect noise or variability in RT data.

A researcher might want to use a proficiency metric to filter out the less proficient members of these samples and compare results between more proficient groups. Generally, a

fine-grained measure such as MINT affords matching of groups either at an individual level (by ensuring that each bilingual in one group has a bilingual in the other group with approximately the same MINT score) or at a group level (ensuring that the mean MINT score for one bilingual group is the same as the mean MINT score for the other). Such matching is a preferable strategy for ensuring similar proficiency between groups. Here, we instead filtered groups based on a threshold MINT score (e.g. people who score 75% or better on the MINT), so as to use a procedure that aligns with that used with self-ratings. That is, matching or filtering on the basis of bilinguals reporting a self-rating of 7 corresponds to filtering based on MINT scores at a 75% threshold or higher. Due to the more coarse nature of the self-rating scale it likely cannot be made more fine grained, even if lengthened or expanded with more elaborate interviewing, due to the imprecision of the introspective process that yields self-ratings.

In this simulation, we used only participants who rated themselves as 7 out of 7 in Spanish or Chinese speaking proficiency (72.6% of Chinese-English bilinguals and 39.5% of Spanish-English bilinguals, n = 162 and 392 respectively). We then pulled a random sample of 40 participants from each of the two groups. The means of the dummy RTs for the Spanish-English bilinguals was 691.4 (*SD* = 120.6) and for the Chinese-English bilinguals was 616.9 (108.7) and these are in fact significantly different ($t = 2.90$, $p < .01$). Alternatively, when we take a sample of the same size based on MINT scores, using only participants who scored at least 75% on the MINT (this threshold was chosen because it also represents 75.8% of Chinese-English and 33.7% of Spanish-English bilinguals in our sample, n = 160 and 280 respectively), we get means of 639.7 (110.7) for the Spanish-English bilinguals and 603.4 (104.1), which are not significantly different ($t = 1.51$, $p = .13$).

Repeating this simulation 10,000 times showed that when self-rated proficiency is used to (hypothetically) match participants, 89.1% of the samples produced significant between group differences. However, when the MINT with a threshold of 75% was used, 68.3% of the simulations yield significant differences between groups. Furthermore, when a MINT threshold of 88% – a number representing 53.1% of Chinese-English bilinguals and 5.5% of Spanish-English bilinguals (n = 112 and 46 respectively), and a much more stringent filter for bilinguals to be considered highly proficient – is used, only 4.9% of the results show significant differences (which, given that the alpha for this statistical test is set at .05, falls within an acceptable range).

All of the effects from both the self-ratings sample and the less stringent MINT samples were in the same direction such that the Chinese-English bilinguals had faster RTs. Our sample of bilinguals, though considerably larger than a typical between-group sample size, still showed a typical difference in data collection. Chinese-English bilinguals, after being filtered for having native-like proficiency, differed systematically from their Spanish-English counterparts. For example, of the participants who rate themselves as 7 out of 7 in other-language speaking, a higher percentage of these Chinese-English participants are Chinese-dominant (62.9%) as compared to the Spanish-dominant Spanish-English bilinguals (34.9%). Whenever a sample of 40 participants from each group is taken, it is more likely to be comparing higher proficiency, Chinese-dominant bilinguals to English-dominant Spanish-English bilinguals than participants who are actually matched in proficiency. This is also true for the 75% MINT threshold samples; 55.0% of Chinese-English bilinguals in this group were Chinese-dominant, whereas 31.4% of Spanish-English bilinguals were Spanish-dominant. Filters as lax as these allow other differences between bilinguals to skew results in misleading directions. In the more stringent MINT sample,

however, 96.4% of Chinese-English bilinguals were Chinese-dominant and 76.1% of Spanish-English bilinguals were Spanish-dominant.

This simulation shows that using an objective measure like the MINT, even at a lower threshold that nearly matches the percentage of a population that would rate themselves perfectly in other-language speaking, can reduce the number of significant differences between seemingly matched groups from 89.1% to 68.3% for our simple dummy variable. Even with more lenient filters, an objective measure of proficiency like the MINT provides a snapshot of language proficiency that is much less likely to suggest group differences when there are none, and the more stringent the filter that is applied, the better the snapshot is that results.

General Discussion

The analyses presented here revealed five primary important differences in how different types of bilinguals rate their proficiency. First, self-ratings of language proficiency varied across bilinguals of different language combinations. Second, differences remained even after organizing populations into discrete groups based on language-combination and dominance (Analysis 3.1), or along a continuous measure of language dominance (Analysis 3.2). Third, Chinese speakers recruited from different linguistic backgrounds showed differences suggesting that different recruitment criteria can create differences in the reference frame bilinguals use to judge proficiency (Analysis 4). Fourth, between-population differences remained significant even after separately considering how well bilinguals could rate their own proficiency level in their non-dominant versus dominant languages separately (Analysis 5). Finally, we simulated a typical reaction time study comparing two language populations and demonstrated that these shortcomings in the self-ratings could potentially lead researchers to draw incorrect conclusions. These analyses are summarized in Table 17.

<Insert Table 17 about here>

As mentioned, simple comparisons between self-ratings and the MINT between populations revealed that Chinese-English bilinguals score more extremely at either end of the self-rating scale than Spanish-English bilinguals. It might have seemed that this difference could occur because of shortcomings of the MINT (based, for example, on the specific items used), but three main findings argue against this possibility. First, there were significant between-group differences in multiple analyses of English MINT scores at any given self-rating except for the highest, even though the test in English was identical for both speakers of Spanish and Chinese (see Tables 2, 3, and Figure 1). Second, there were significant differences between bilinguals dominant in one versus the other language even within bilinguals of the same language combination, and third, these differences were in opposite directions at the two ends of the scale (see Tables 7, 8 and Figure 3). Considerable within-population differences cannot be explained by an ineffectiveness of the MINT to capture language or cultural differences. The Spanish-dominant Spanish-English bilinguals came from similar cultural and geographic backgrounds as the English-dominant Spanish-English bilinguals. The majority in both cases (68.8% of Spanish-dominant and 90.2% of the English-dominant Spanish-English bilinguals) were born in the USA (with Mexico being the second highest demographic representing 22.7% and 6.98% of the respective populations). Finally, the MINT patterned similarly between languages when compared to the Oral Proficiency Interview scores, suggesting that any differences in the other analyses come from differences in self-ratings, and not a problem with the MINT itself.

In the third analysis, we found within-population differences based on language dominance - other-language dominant bilinguals named fewer pictures in English than their English-dominant peers, even at the same self-ratings (see Table 8 and Figure 3). This suggests

that even groups recruited within the same population may differ in their self-assessment of language proficiency. In addition to these within-population differences, we found problematic differences between populations. Specifically, English-dominant Chinese-English bilinguals scored lower than both their Chinese-dominant Chinese-English peers, but also lower than English-dominant Spanish-English bilinguals at the same self-rating. One possible explanation for this pattern of results is that different participants have different frames of reference that they use to evaluate their language proficiency. For instance, bilinguals recruited for an experiment from a population of Spanish-English bilinguals in San Diego may rate their proficiency a 5 or 6 out of 7 in English speaking proficiency, judging that they are relatively less fluent than their peers at UCSD. They may not, however, judge themselves against highly Spanish-dominant Spanish-English speakers from Mexico, Spanish heritage speakers in the northeast United States (where environmental exposure is less compared to southern California), or any other nonnative English speaker. The MINT, and indeed proficiency as a metric in cognitive testing, is not biased by participant reference frame or bilingual subpopulation.

We explored this possibility by comparing three separate populations of Chinese speakers: a group exposed to Chinese in their home growing up, the group of bilinguals in Analyses 1 and 3, recruited only for native knowledge of both languages, and a group of Chinese speaking students that were recruited for having relatively low English proficiency. We compared their self-ratings to MINT scores (see Tables 12, 13 and Figure 4) in both English and Chinese and found that the relationship between self-ratings and MINT scores differed significantly by recruitment group, particularly in Chinese, even at the same university and even within the same language, when both of the languages of the bilingual population are the same. This suggests that internal reference frame can vary based on the bilingual's own subpopulation.

Though we show here that self-ratings may vary by internal reference frame, they may be more reliable within a bilingual's own system (e.g., a bilingual may know that their English is better than their Spanish, and therefore give it a higher rating). Consistent with this view, when bilinguals' responses were separated into how they rated their own dominant and non-dominant languages (instead of by English, Spanish, or Chinese), subjective measures performed closer to objective measures (Analysis 5). There was still a significant crossover interaction in the dominant language (that patterns the same as in Analyses 1 and 3) such that Chinese-English bilinguals had better MINT scores at higher ends of the scale and worse at lower scores (though the differences were a bit smaller at the lower end). This interaction was absent in the non-dominant language, however, there was still a substantial main effect such that Chinese-English bilinguals performed better in their non-dominant language than Spanish-English bilinguals did in their non-dominant language at any given self-rating. Though different from the cross-over interactions observed in the other analyses, it arguably reveals an equally problematic case in which any comparison made between two populations at a certain self-rating would still lead to erroneous conclusions about the relationship between that population's proficiency and the effect of interest.

Though bilinguals fared better in self-assessment of which language is dominant, a major exception was found in those bilinguals who rated themselves as balanced. 77.8% of self-assessed balanced bilinguals were actually more dominant in one language or the other (based on a 5% margin in MINT scores used from Gollan et al. 2012), in line with previous work showing that bilinguals are rarely truly balanced in both languages (Grosjean, 1982). This demonstrates another way in which bilinguals' self-assessment of their own proficiency levels in each language is problematic.

In these analyses, we primarily used a different approach to assessing dominance – averaging all four modalities within a language before comparing scores to determine balanced status. This offers a more nuanced self-rating of language proficiency. One might argue that because the MINT is a measure of speaking proficiency, and because we used self-rated speaking as the independent variable in our critical analyses, this measure alone should determine dominance. Exploring that possibility revealed some potential problems with this approach. First, the number of bilinguals that would be classified as "balanced" increased (from 12 to 39 for balanced Chinese-English bilinguals, and from 162 to 374 balanced Spanish-English bilinguals). However, if self-rated speaking was indeed a better indication of dominance in MINT scores, this number should instead decrease, as MINT scores indicate that these bilinguals were significantly better at speaking in one of their languages. Additionally, two models had factors that became nonsignificant (due partially to the increased number of bilinguals that were classified as balanced). In Analysis 3.1 the interaction between self-ratings and language dominance that indicated that dominance groups differed in other-language MINT score based on their dominance group and self-rating became nonsignificant. Likewise, the interaction between language dominance and language combination that showed that subjects scored differently in their other-language MINT based on their subgroup determined with respect to their language dominance and language combination (see Table 7 and Figure 3a) became nonsignificant. Additionally, in Analysis 5, there was an interaction in the dominant language condition that showed a crossover between population in predicting dominant language MINT score (see Table 14 and Figure 5a) that became nonsignificant. While these interactions suggested different ratings at opposite ends of the scale, even the main effects showed significant, systematic bias in the same direction such that one population had higher MINT

scores relative to other given the same ratings at all points on the scale. Furthermore, the Edinburgh dominance measure also showed that language dominance, language combination and self ratings significantly impacted MINT scores (Analysis 3.2). These differences therefore do not alter the conclusions drawn – different populations and dominance groups rate themselves differently, no matter how the data are organized, and this can provide misleading results.

As we have seen, between-participant self-ratings can become misleading in many cases – especially when bilinguals of different language combinations, cultures, or dominance profiles are treated as if they represent one homogenous population (at least with respect to how they provide self-ratings). Significant correlations between two measures like self-ratings and objective proficiency reveal that the two measures pattern together, but do not imply that the two will pattern sufficiently closely in all comparisons and for all purposes. Marian et al. (2007) reported that self-ratings (paired with language use questions in the same factor of a factor analysis) can account for about 25% of the variance in objective measures of proficiency, which translates to a correlation of about .5. Though this shows that the two measures are related, it leaves enough room for divergence between the self-ratings and actual proficiency, which could lead to problematic conclusions. Self-rated proficiency measures are common in experiments with bilinguals and are of course better than no measure of proficiency at all. However, the results that come from using self-ratings can be misleading in many cases; the simulation showed that the MINT – or perhaps other comparable objective measures – can likely better account for differences (or non-differences) between populations than self-ratings, and will therefore lead to greater accuracy in interpretation of results and improved consistency in results across experiments carried out by different experimenters with different language populations in different settings.

These analyses have demonstrated breakdowns in seemingly straightforward assumptions commonly made in bilingual research and how use of objective measures could improve measurement and consistency between studies of different types of bilinguals. Frame of reference is a widely studied topic that could benefit bilingualism researchers looking at population level differences in self-rating. However, for studies that need a reliable metric of language proficiency, objective measures are the better choice. Of course, objective measures are not direct quantifications of language proficiency and can themselves be problematic, particularly when not designed specifically to measure proficiency in the target languages (e.g., the Boston Naming Test was developed for English speakers but is often used to assess proficiency in bilinguals of various language combinations; for examples see Allegri et al., 1997; Kohnert et al., 1998; Gollan et al., 2007; Patricacou et al., 2007; Silverberg & Samuel, 2004). Experimental and clinical psychologists are tasked with finding the most valid behavioral measures, but we have suggested that self-ratings are systematically biased and flawed – and should not be relied upon whenever tried and true objective measures are available, and should be interpreted with great caution when objectives measures are not available. Proficiency comparisons between language populations and between levels of experience or dominance within language combinations can be misleading; when interpreting self-ratings, one person's 7 might be more like someone else's 5.

References

Allegri, R. F., Villavicencio, A. F., Taragano, F. E., Rymberg, S., Mangone, C. A., & Baumann, D. (1997). Spanish Boston naming test norms. *The Clinical Neuropsychologist*, *11*(4), 416-420.

Dunn, A. L., & Tree, J. E. F. (2009). A quick, gradient bilingual dominance scale. *Bilingualism: Language and Cognition*, *12*(03), 273-289.

Gollan, T. H., Fennema-Notestine, C., Montoya, R. I., & Jernigan, T. L. (2007). The bilingual effect on Boston Naming Test performance. *Journal of the International Neuropsychological Society*, *13*(02), 197-208.

Gollan, T. H., Starr, J., & Ferreira, V. S. (2015). More than use it or lose it: The number-of-speakers effect on heritage language proficiency. *Psychonomic bulletin & review*, *22*(1), 147-155.

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*(3), 594.

Grosjean, F. (1982). *Life with two languages: An introduction to bilingualism*. Harvard University Press.

Grosjean, F. (1998). Studying bilinguals: Methodological and conceptual issues. *Bilingualism: Language and cognition*, *1*(2), 131-149.

Hulstijn, J. H. (2012). The construct of language proficiency in the study of bilingualism from a cognitive perspective. *Bilingualism: Language and Cognition*, *15*(2), 422-433.

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*, *19*(03), 272-283.

Kaplan, E. F., H. Goodglass, and S. Weintraub. "The Boston naming test. 2nd." *Philadelphia: Lea & Febiger* (1983).

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*(3), 422-440.

Li, P., Sepanski, S., & Zhao, X. (2006). Language history questionnaire: A web-based interface for bilingual research. *Behavior research methods*, *38*(2), 202-210.

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*(4), 940-967.

Patricacou, A., Psallida, E., Pring, T., & Dipper, L. (2007). The Boston Naming Test in Greek: Normative data and the effects of age and education on naming. *Aphasiology*, *21*(12), 1157-1170.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Schwarz, N. (1999). Self-reports: how the questions shape the answers. *American psychologist*, *54*(2), 93.

Sheng, L., Lu, Y., & Gollan, T. H. (2014). Assessing language dominance in Mandarin–English bilinguals: Convergence and divergence between subjective and objective measures. *Bilingualism: Language and Cognition*, *17*(02), 364-383.

Silva-Corvalán, C., & Treffers-Daller, J. (2016). *Language dominance in bilinguals: Issues of measurement and operationalization*. Cambridge University Press.

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*(3), 381-398.

Tao, L., Taft, M., & Gollan, T. H. (2015). The bilingual switching advantage: Sometimes related to bilingual proficiency, sometimes not. *Journal of the International Neuropsychological Society*, *21*(07), 531-544.

Zell, E., & Krizan, Z. (2014). Do people have insight into their abilities? A metasynthesis. *Perspectives on Psychological Science*, *9*(2), 111-125.

[1] Two interactions in Table 7 (Figure 3a) became nonsignificant when using only self-rated speaking (as opposed to the average of all four self-ratings) to determine dominance – the interaction between self-ratings and dominance and the three-way interaction between those factors and language population. Additionally, the percentage of people who classify themselves as balanced when their MINT scores differ (beyond a strict margin of 5%) increases from 77.8% to 87.0%, despite the fact that this classification is specified to the modality of speaking, theoretically giving bilinguals a better chance of self-assessment. Though using self-rated speaking as a determiner of dominance is less stringent and more specific to the MINT (in which the modality is speaking), the changes in model significance were likely the result of the removal of 239 participants, as well as the increase in misclassification of dominance; the pattern of results remained the same, and all remaining main effects and the interaction between language self-ratings and language combination remained significant. Any main effects are still problematic for interpretation– two populations that respond significantly differently in their self-ratings may lead to erroneous conclusions.

[2] This interaction was no longer significant when redoing the analysis while relying only on self-rated speaking to classify bilinguals into groups. Instead, with this change, Spanish-English bilinguals scored higher than the Chinese-English bilinguals at any given self-rating (exactly the opposite pattern relative to what is reported for the non-dominant language; see Figure 5b). Additionally, 239 bilinguals had to be excluded from the analysis because they became "balanced" when relying only self-rated speaking (instead of the average of ratings for all four modalities). However, these differences do not alter the interpretation of results – bilinguals in these two populations behave significantly differently when self-assessing both their dominant and non-dominant languages.

[3] Note that all statistics were done both with regression (reported), as well as linear mixed effect models, treating the experiment each subject originated from as a random variable. No significant differences in coefficient estimation statistics arose as a result of this difference (though model comparisons, not coefficient estimation statistics, are reported here).

[4] Note that all reported regression coefficients are unstandardized.

[5] Note that the correlation using all four modalities averaged between English rating minus other-language rating and English rating minus other-language MINT score in Chinese speakers is highest in part due to the fact that the sample of Chinese-English bilinguals was more balanced in dominance compared to the Spanish-English bilinguals – taking the absolute value of the numbers in this correlation reduced the correlation from .87 to .50.

Table 1a. *Participant characteristics of Spanish-English bilinguals from Analyses 1,3 and 5.*

		English-Dominant (n = 702)			Spanish-Dominant (n = 128)			Balanced (n = 162)		
		M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
	Age	20.32	2.20	18-35	20.73	2.52	18-33	20.36	1.91	18-28
	% Female	76.78	n/a	n/a	64.84	n/a	n/a	74.69	n/a	n/a
	Education	14.01	1.39	12-20	14.22	1.53	12-19	14.16	1.68	12-24
	Primary parent Education	11.16	4.22	0-21	11.26	4.11	2-21	11.33	4.24	0-21
	Secondary parent Education	10.65	4.58	0-60	11.21	4.18	2-21	10.59	4.61	0-21
	% English use daily Currently	85.19	11.91	20-100	67.80	20.97	15-100	79.71	15.30	20-100
	% English use daily Growing up	59.68	15.10	5-95	37.04	19.67	0-90	49.01	15.76	10-90
English	Age 1st Exposure	3.03	2.42	0-13	5.98	3.28	0-18	3.97	2.78	0-13
	Self-rated Speaking	6.74	0.55	3-7	5.58	0.74	3-7	6.55	0.72	3-7
	Self-rated Reading	6.74	0.53	4-7	5.71	0.84	3-7	6.58	0.73	2-7
	Self-rated Writing	6.62	0.66	4-7	5.39	0.97	2-7	6.47	0.83	2-7
	Self-rated Listening	6.83	0.44	4-7	6.13	0.73	4-7	6.72	0.55	5-7
	MINT	61.14	3.43	32-68	56.45	4.73	43-67	59.57	3.34	48-67
Spanish	Age 1st Exposure	0.48	1.22	0-12	0.37	0.89	0-6	0.42	0.91	0-5
	Self-rated Speaking	5.71	1.05	1-7	6.68	0.59	5-7	6.57	0.76	3-7
	Self-rated Reading	5.56	1.02	3-7	6.38	0.84	3-7	6.54	0.78	2-7
	Self-rated Writing	4.84	1.06	1-7	6.10	1.03	3-7	6.35	0.96	2-7
	Self-rated Listening	6.35	0.86	1-7	6.80	0.44	5-7	6.75	0.55	4-7
	MINT	44.48	9.17	10-67	53.82	7.38	27-68	50.77	7.36	32-64

Table 1b. *Participant characteristics of Chinese-English bilinguals from Analyses 1,3 and 5.*

		English-Dominant (n = 72)			Mandarin-Dominant (n = 139)			Balanced (n = 12)		
		M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
	Age	20.16	1.50	18-25	20.29	1.77	18-28	20.17	0.72	19-22
	% Female	62.50	n/a	n/a	78.41	n/a	n/a	66.67	n/a	n/a
	Education	13.61	2.03	0-16	13.63	1.18	12-17	13.75	1.14	12-16
	Primary parent Education	15.50	3.66	2-21	16.25	2.73	3-21	14.75	2.45	11-18
	Secondary parent Education	14.98	4.16	0-21	16.32	2.60	3-21	14.81	3.06	12-21
	% English use daily Currently	87.72	11.63	50-100	59.28	22.00	10-100	77.42	16.05	51-98
	% English use daily Growing up	58.13	16.53	10-90	20.32	16.94	0-90	45.08	15.26	25-80
English	Age 1st Exposure	3.23	2.95	0-11	6.39	2.87	0-16	4.25	3.65	0-13
	Self-rated Speaking	6.86	0.48	4-7	5.27	0.90	2-7	6.67	0.65	5-7
	Self-rated Reading	6.81	0.55	4-7	5.36	0.79	3-7	6.50	0.80	5-7
	Self-rated Writing	6.64	0.74	4-7	5.10	0.84	3-7	6.42	0.90	5-7
	Self-rated Listening	6.89	0.36	5-7	5.64	0.82	4-7	6.58	0.67	5-7
	MINT	63.51	3.40	53-68	50.35	5.68	35-66	56.67	7.29	36-63
Mandarin	Age 1st Exposure	2.08	2.86	0-13	1.27	2.14	0-19	1.17	1.99	0-7
	Self-rated Speaking	5.77	1.05	3-7	6.94	0.29	5-7	6.58	0.67	5-7
	Self-rated Reading	4.18	1.57	1-7	6.91	0.45	3-7	6.42	0.79	5-7
	Self-rated Writing	3.34	1.57	1-6	6.74	0.82	2-7	6.17	1.11	4-7
	Self-rated Listening	6.00	0.94	4-7	6.97	0.17	6-7	6.67	0.65	5-7
	MINT	43.75	10.25	22-62	60.74	3.34	35-66	52.33	6.83	35-60

Table 2. *Regression of other-language MINT scores on to subjective self-rating speaking ability and language combination for Analysis 1, shown in Figure 1a.*[3]

	Coefficient		Test statistic			
	B[4]	*SE*	SSE	MSE	*F*	*p*
Self-Rated Speaking	8.81	0.66	27688	27688	396.28	<.001
Language combination	28.94	4.61	5573	5573	79.76	<.001
Interaction	-5.35	0.71	4021	4021	57.55	<.001

adj. $R^2 = 0.30$

Table 3. *Regression of English MINT onto subjective self-rating speaking ability and language combination for Analysis 1, shown in Figure 1b.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-Rated Speaking	4.96	0.26	10307	10307	591.15	<.001
Language combination	21.51	1.97	1718	1718	98.54	<.001
Interaction	-3.00	0.32	1551	1551	88.94	<.001

adj. $R^2 = 0.39$

Table 4. *Participant characteristics for Analysis 2, adapted from Gollan et al. (2012) and Sheng et al. (2014).* See original publications for full participant characteristics. Note that Self-Rated Speaking is out of a possible 10 rather than 7 and MINT is out of a possible 1.0 rather than 68.

		Spanish-English (n = 52)			Chinese-English (n = 62)		
		M	*SD*	Range	*M*	*SD*	Range
	Age	20.77	2.93	18-36	19.83	1.29	18-24
English	Self-rated Speaking	9.22	1.03	6-10	8.77	1.18	6-10
	Oral Proficiency	8.75	0.97	6.5-10	8.82	1.09	5-10
	MINT	0.89	0.05	0.75-0.97	0.89	0.08	0.56-1
Other Language	Self-rated Speaking	8.35	1.25	5-10	7.67	1.42	5-10
	Oral Proficiency	7.76	1.12	6-10	7.26	1.79	3-10
	MINT	0.73	0.13	0.38-0.93	0.69	0.19	0.12-0.93

Table 5. *Regression of other-language MINT score onto OPI score and language combination for Analysis 2, shown in Figure 2a.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
OPI Score	0.08	0.01	1.56	1.56	116.84	<.001
Language combination	0.15	0.13	0.00	0.00	0.02	0.89
Interaction	-0.02	0.02	0.02	0.02	1.50	0.22

adj. R^2 = 0.51

Table 6. *Regression of English MINT onto OPI score and language combination for Analysis 2, shown in Figure 2, shown in Figure 2b.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
OPI Score	0.04	0.01	0.12	0.12	32.90	<.001
Language combination	0.13	0.10	0.01	0.01	1.49	0.22
Interaction	-0.02	0.01	0.01	0.01	2.20	0.14

adj. R^2 = 0.23

Table 7. *Regression of other-language MINT score onto subjective self-rated speaking proficiency, language combination and categorical language dominance for Analysis 3, shown in Figure 3a.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-rated Speaking	4.80	0.90	26542	26542	411.31	<.001
Language combination	14.49	5.55	5803	5803	89.92	<.001
Language Dominance	-0.48	17.26	9705	9705	150.40	<.001
Self-rated Speaking: Language combination	-2.34	0.95	1767	1767	27.38	<.001
Self-rated Speaking: Language Dominance	1.70	2.53	408	408	6.34	.012
Language combination: Language Dominance	-10.71	19.10	338	338	5.24	.022
Self-rated Speaking: Language combination: Language Dominance	0.99	2.81	8	8	0.12	.72

adj. R^2 = 0.40

Table 8. *Regression of English MINT onto subjective self-rated speaking proficiency, language combination and categorical language dominance for Analysis 3, shown in Figure 3b.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-rated Speaking	3.08	0.95	10960	10960	737.08	<.001
Language combination	11.13	6.75	1444	1444	97.09	<.001
Language Dominance	-0.99	6.79	2381	2381	160.15	<.001
Self-rated Speaking: Language combination	-1.95	0.98	1107	1107	74.44	<.001
Self-rated Speaking: Language Dominance	-1.38	1.01	74	74	4.97	.026
Language combination: Language Dominance	-7.16	7.50	832	832	55.95	<.001
Self-rated Speaking: Language combination: Language Dominance	2.23	1.15	56	56	3.79	0.052

adj. R^2 = 0.52

Table 9. *Regression of other-language MINT score onto subjective self-rated speaking proficiency, language combination and Edinburgh language dominance for Analysis 3.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-rated Speaking	0.88	0.44	27688	27688	3137.80	<.001
Language combination	3.77	3.10	5573	5573	631.53	<.001
Language Dominance	-81.48	10.13	76473	76473	8666.63	<.001
Self-rated Speaking: Language combination	-0.33	0.46	168	168	19.04	<.001
Self-rated Speaking: Language Dominance	4.44	1.53	27	27	3.06	0.081
Language combination: Language Dominance	19.68	10.98	1174	1174	133.05	<.001
Self-rated Speaking: Language combination: Language Dominance	-6.76	1.70	140	140	15.84	<.001

adj. R^2 = 0.91

Table 10. *Regression of English MINT onto subjective self-rated speaking proficiency, language combination and Edinburgh language dominance for Analysis 3.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-rated Speaking	0.53	0.32	10307	10307	766.03	<.001
Language combination	-7.59	2.34	1718	1718	127.69	<.001
Language Dominance	118.46	14.07	5507	5507	208.93	<.001
Self-rated Speaking: Language combination	1.50	0.37	704	704	87.94	<.001
Self-rated Speaking: Language Dominance	-12.45	2.12	678	678	60.43	<.001
Language combination: Language Dominance	-63.73	16.53	1522	1522	108.62	<.001
Self-rated Speaking: Language combination: Language Dominance	6.01	2.48	69	69	11.60	0.016

adj. R^2 = 0.59

Table 11. *Participant characteristics from Analysis 4*. Note that one experiment did not solicit self-ratings for the categories of reading and writing. Note that Education, and primary/secondary parent education was not available for these studies.

		Recently Immigrated Chinese (n = 144)			Chinese Exposed (n = 90)			Chinese-English (n = 223)		
		M	*SD*	Range	*M*	*SD*	Range	*M*	*SD*	Range
	Age	20.29	2.22	18-34	19.8	1.22	18-23	20.25	1.65	18-28
	% Female	66.67	n/a	n/a	74.4	n/a	n/a	72.65	n/a	n/a
	% English use daily Currently	52.43	21.31	1-100	93.96	8.64	60-100	69.44	23.09	10-100
	% English use daily Growing up	20.07	18.72	0-100	72.28	17.31	15-100	33.87	24.29	0-90
	Age Moved to U.S.	15.48	5.39	0-34	-	-	-	-	-	-
English	Age 1st Exposed	7.06	3.04	0-16	0.97	1.67	0-6	5.27	3.28	0-16
	Self-rated Speaking	4.99	1.36	1-7	6.96	0.18	6-7	5.86	1.09	2-7
	Self-rated Reading	-	-	-	6.9	0.3	6-7	5.89	0.99	3-7
	Self-rated Writing	-	-	-	6.84	0.48	5-7	5.67	1.09	3-7
	Self-rated Listening	5.43	1.12	2-7	6.92	0.31	5-7	6.09	0.91	4-7
	MINT	49.7	7.91	9-64	64.35	2.45	57-68	54.94	7.97	35-68
Other Language	Age 1st Exposed	0.22	0.92	0-6	1.24	2.62	0-12	1.53	2.41	0-19
	Self-rated Speaking	6.44	1.03	2-7	4.28	1.57	1-7	6.55	0.85	3-7
	Self-rated Reading	-	-	-	2.8	1.4	1-7	6.00	1.59	1-7
	Self-rated Writing	-	-	-	2.52	1.2	1-6	5.62	1.94	1-7
	Self-rated Listening	6.59	0.86	2-7	4.72	1.66	1-7	6.64	0.72	4-7
	MINT	58.66	4.04	36-65	26.62	16.04	0-55	54.81	10.24	22-66

Table 12. *Regression of Chinese MINT onto subjective self-rated speaking and bilingual type for Analysis 4, shown in Figure 4a.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-rated Speaking	7.94	0.48	82729	82729	1718.50	<.001
Bilingual Type – Chinese-English	4.35	4.26	7060	3530	73.33	<.001
Bilingual Type – Chinese Immigrated	41.17	5.55	-	-	-	-
Self-rated Speaking: Bilingual Type – Chinese-English	0.94	0.73	1502	751	15.60	<.001
Self-rated Speaking: Bilingual Type – Chiniese Immigrated	-4.15	0.90	-	-	-	-

adj. R^2 = 0.81

Table 13. *Regression of English MINT on subjective self-rated speaking and bilingual type for Analysis 4, shown in Figure 4b.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-rated Speaking	0.37	3.17	21125	21125	725.91	<.001
Bilingual Type – Chinese-English	-35.74	22.17	986	493	16.95	<.001
Bilingual Type – Chinese Immigrated	-32.77	22.15	-	-	-	-
Self-rated Speaking: Bilingual Type – Chinese-English	4.56	3.19	85	42	1.45	.24
Self-rated Speaking: Bilingual Type – Chinese Immigrated	4.06	3.19	-	-	-	-

adj. R^2 = 0.62

Table 14. *Regression of dominant-language MINT on subjective self-rated speaking and language combination for Analysis 5, shown in Figure 5a.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-Rated Speaking	4.05	0.87	1601	1601	74.80	<.001
Language combination	12.89	6.31	268	268	12.50	<.001
Interaction	-2.05	0.91	108	108	5.06	.025

adj. R^2 = 0.08

Table 15. *Regression of nondominant-language MINT onto subjective self-rated speaking and language combination for Analysis 5, shown in Figure 5b.*

	Coefficient		Test statistic			
	B	*SE*	SSE	MSE	*F*	*p*
Self-Rated Speaking	2.00	0.64	4667	4667	56.30	<.001
Langauge Group	-3.45	3.97	822	822	9.82	.002
Interaction	0.23	0.71	8	8	0.10	.75

adj. R^2 = 0.06

Table 16. *Correlations between self-rated proficiency scores or their difference and MINT scores.* All correlations were significant at $p < .001$. Participant information is listed in Table 1.

		All Four Modalities Averaged		Self-rated Speaking	
		Spanish speakers	Chinese speakers	Spanish speakers	Chinese speakers
Self-rating separated by language	English MINT	.34	.69	.36	.67
	Spanish/Chinese MINT	.39	.82	.39	.73
Self-rating separated by dominance	Dominant MINT	.25	.39	.12	.31
	Non-dominant MINT	.33	.26	.55	.44
English rating minus other-language rating	English MINT minus other-language MINT	.55	. 87[5]	.53	.81
Dominant rating minus non-dominant rating	Dominant MINT minus non-dominant MINT	.40	.27	.68	.50

Table 17. *Summary of analysis outcomes*

	Outcome	Details
Analysis 1	The relationship between self-ratings and an objective test of picture-naming (MINT) differs between bilingual populations.	Figure 1. Chinese-English bilinguals with high self-ratings had higher objectively measured Chinese ability than Spanish-English bilinguals with the same self-rating for Spanish. Conversely, Chinese-English bilinguals with low self-ratings for Chinese, had lower objectively measured Chinese ability than Spanish-English bilinguals with the same self-rating. The difference at the low end was apparent for English ratings as well.
Analysis 2	The relationship between two objective measures (the MINT and Oral Proficiency Interviews) does not differ significantly between bilingual populations.	Figure 2. The correlation between OPI and MINT scores was similar regardless of bilingual-language combination and regardless of language (Chinese/Spanish or English).
Analysis 3	The relationship between self-ratings and the MINT differs based on language dominance, even after factoring in bilingual population.	Figure 3. Self-rated English-dominant Spanish-English bilinguals named more pictures in their other language at any given point on the self-rated proficiency scale than English-dominant Chinese-English bilinguals. Furthermore, self-rated other-language dominant bilinguals name more other-language pictures than their English dominant peers, even at the same self-rating, regardless of language population. This was true for both English and other-language MINT scores.
Analysis 4	The relationship between self-ratings and MINT differ even within bilinguals of the same language combination (recently immigrated, Chinese-English bilingual, or exposed Heritage speakers).	Figure 4. Chinese-English bilinguals recruited for having recently immigrated named more pictures in Chinese at any given self-rating than Chinese-English bilinguals recruited only for speaking both languages, who in turn named more than those recruited for being exposed to Chinese.

Analysis 5	Grouping data by dominant and non-dominant languages still reveal problematic differences between self-ratings and MINT score.	Figure 5. Chinese-English bilinguals scored higher at high ratings than their Spanish-English peers in whichever language's MINT test they considered to be their dominant language, and lower at lower ratings (similar to Analysis 1). They also scored higher in their non-dominant language than their Spanish-English peers at any given self-rating, but there was no interaction between language combination and self-rating.

Figure 1. MINT scores as a function of self-rated proficiency in 992 Spanish-English and 223 Chinese-English bilinguals.

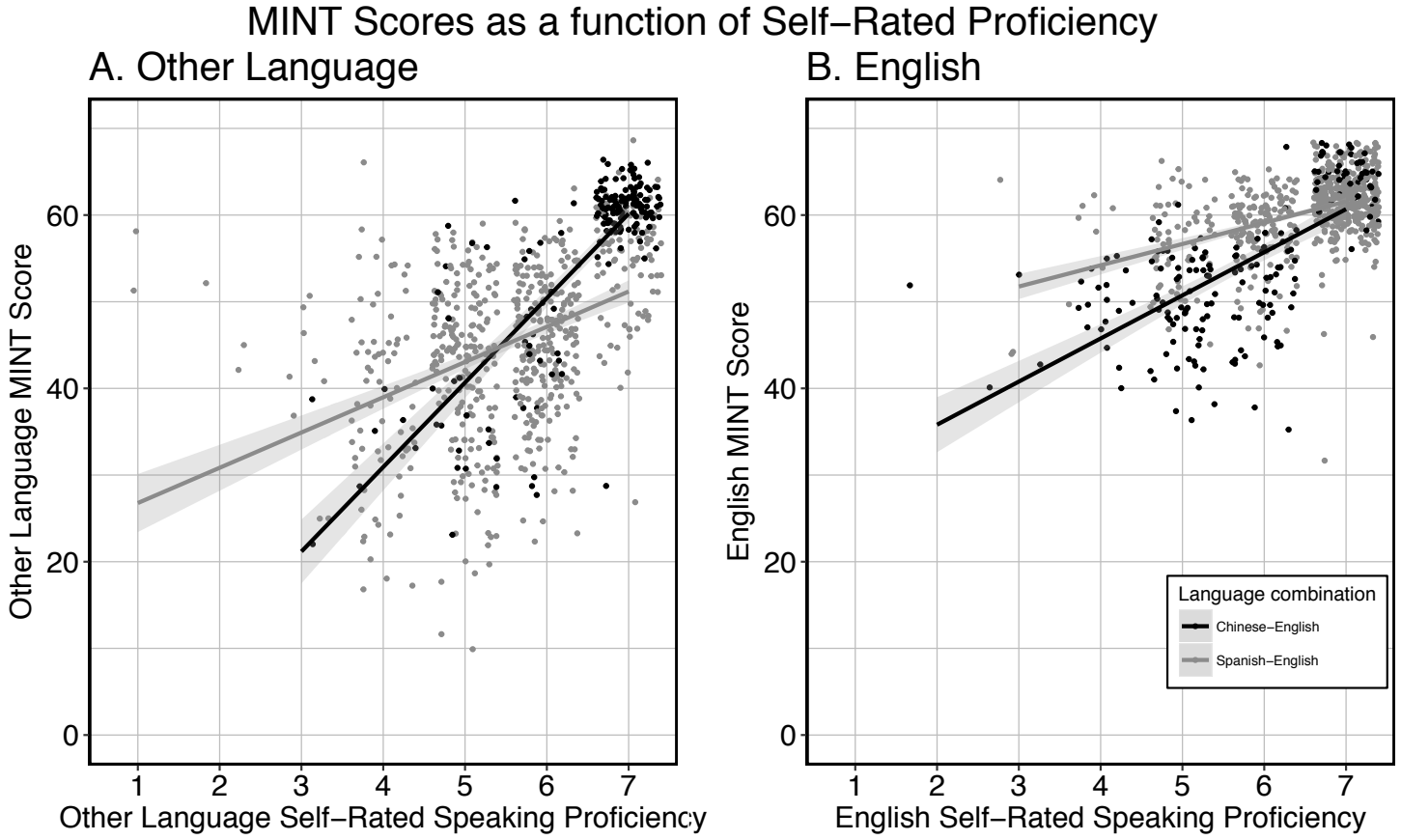

Figure 2. Reanalysis of Gollan et al. (2012) and Sheng et al. (2014) showing MINT scores as a function of Oral Proficiency scores.

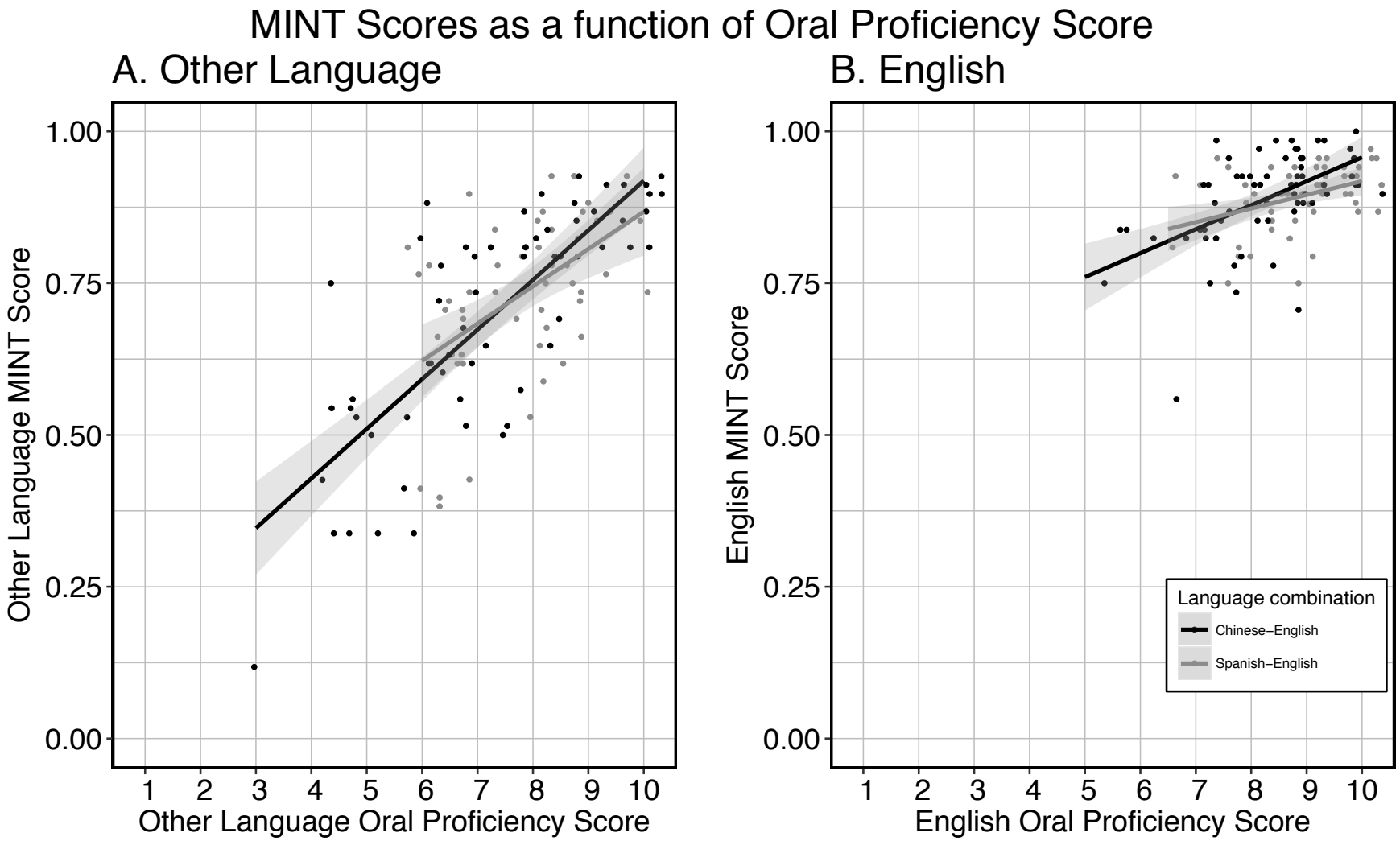

Figure 3. MINT scores as a function of self-rated proficiency and dominance in Spanish-English (black) and Chinese-English (grey). Solid lines represent other-language dominant bilinguals, whereas dashed lines represent English dominance, and alternating dash-dot lines represent balanced bilinguals.

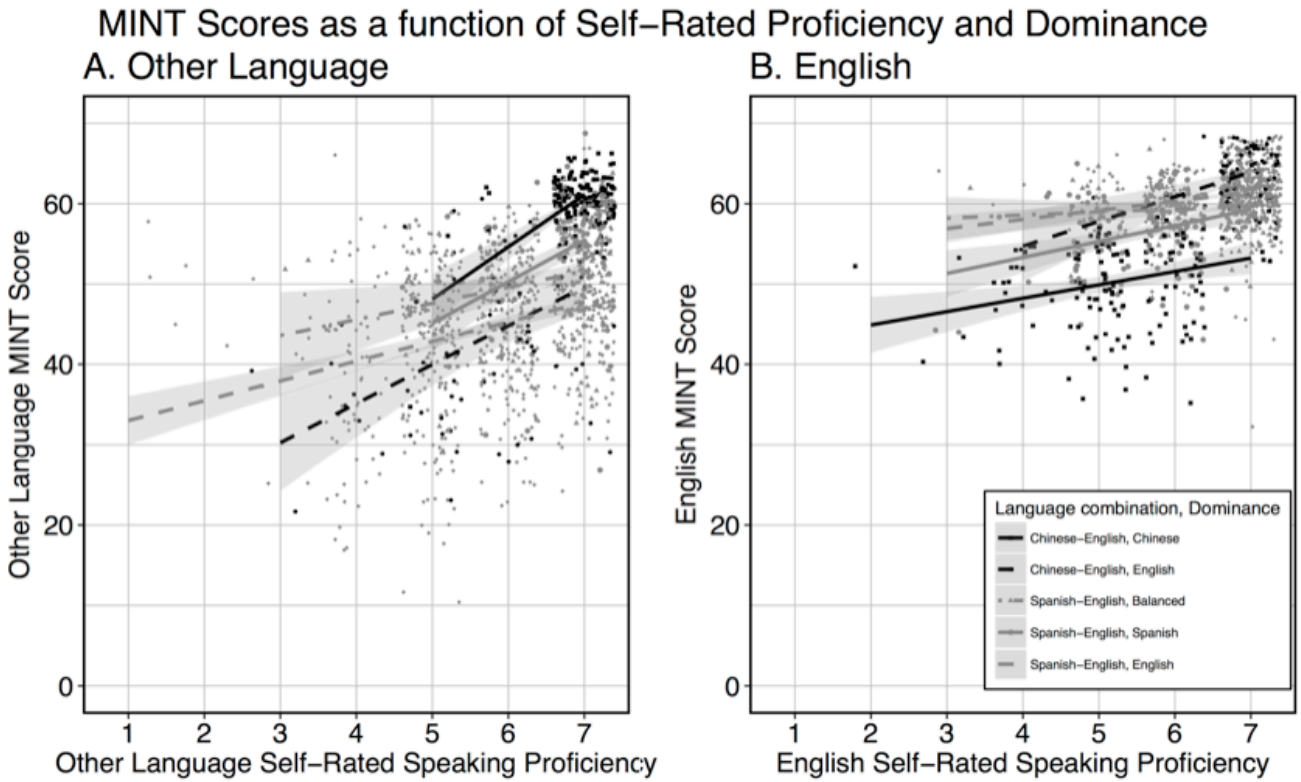

Figure 4. MINT scores as function of self-rated proficiency in three Chinese speaking populations. Chinese exposed speakers are marked with circles, Chinese-English bilinguals with crosses, and recently immigrated Chinese speakers with triangles.

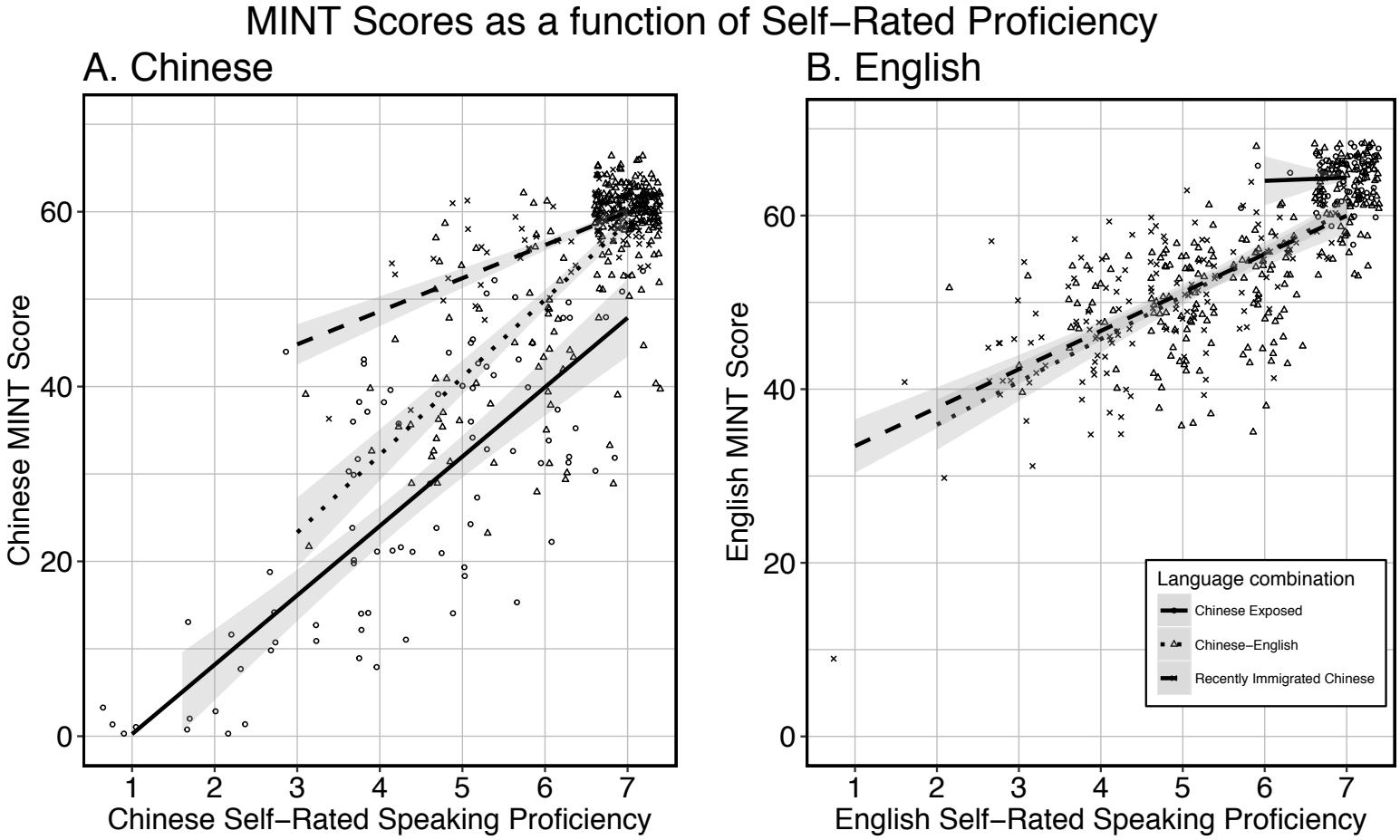

Figure 5. MINT scores as function of self-rated proficiency, collapsed across languages, but separated into Non-Dominant and Dominant Languages, rather than by English or other-language. This plot excludes balanced bilinguals.