Appendix

Scoring of the object naming scale.

We used object naming scores based on IRT, so that scores accounting for DIF could be formed, as explained below. All item response theory IRT scores were computed using PARSCALE 4.1. (41) We used *expectation a posteriori* scoring, which permits finite scores for participants with perfect or zero scores, and the 2-parameter logistic model (42) to analyze the 44 dichotomous items. We compared model fit statistics between the models. Overall fit improved from the unmodified score to the language-only score to the model-adjusted score to the simultaneous score. Overall model fit for the simultaneous score reported by PARSCALE was a χ^2 of 738 on 678 degrees of freedom (P=0.05).

One feature of IRT is that item parameters are invariant (within a linear transformation) across subgroups such as those defined by language, education, or dementia status, as long as IRT model assumptions such as unidimensionality and local independence are met. These assumptions were tested with a single factor confirmatory factor analysis (CFA) using MPLUS,(43) with a weighted least squares estimator. A single factor CFA model fit well, with a confirmatory fit index (CFI) of 0.976, a Tucker-Lewis index (TLI) of 0.988, and a root mean squared error of approximation (RMSEA) of 0.036. The largest modification index was for the correlation between *cloud/nube* and *fog/niebla*. With this empirically guided residual correlation included, CFI was 0.977, TLI was 0.988, and RMSEA was 0.036. We compared the loadings for the single factor model with and without these residual correlations. Standardized factor loadings ranged from 0.44 to 0.85 for both models. Most loadings were unchanged, and all differences in

standardized loadings between the two models were all less than 0.03. We concluded that the object naming scale was sufficiently unidimensional to proceed with IRT analyses.

DIF detection framework.

We have developed an IRT-ordinal logistic regression DIF-detection procedure (9, 33, 44) that we used for all three strategies. We have written a Stata program which will call PARSCALE and run the DIF analyses. It can be downloaded by typing "ssc install difwithpar" in the Stata command window.

We looked for DIF-free items to serve as anchors, which are items that have the same relationship with the underlying cognitive ability in both languages, or, in the simultaneous strategy, in all language-education subgroups. The DIF-free items were used to anchor relationships between test items across the groups so that other items could be compared. Selection of anchor items is an essential aspect of any DIF detection procedure. If items are identified as anchors that nevertheless have DIF, other items may be falsely identified with DIF or falsely declared to be free of DIF; this phenomenon is referred to as spurious DIF.

We outline the general method below, using testing for language DIF as the example. We examined three logistic regression models for each item:

$$\text{logit } p(y=1) = \textit{intercept} + \beta_1 * \hat{\theta} + \beta_2 * \text{language group} + \beta_3 * \hat{\theta} * \text{language group}, \quad (1)$$

logit p(y=1) = *intercept* + β_1*^,θ + β_2* language group,

(2)

logit p(y=1) = *intercept* + β_1*^,θ (3)

In these models, P(y=1) is the probability of naming the object correctly, ^,θ ^,q (theta) is the IRT estimate of object naming ability obtained from PARSCALE, and "language group" is the indicator for Spanish vs. English language use. In model 1, β_3 is the coefficient for the ability-language group interaction term. All DIF analyses were performed using –difwithpar- for Stata (45) (type *ssc install difwithpar* at the Stata prompt to obtain the difwithpar package).

Two types of DIF are identified in the literature. In items with *non-uniform DIF*, demographic interference between ability level and item responses differs at varying levels of object naming ability. In items with *uniform DIF*, the interference is the same across all levels of object naming ability. To detect non-uniform DIF, we compared the log likelihoods of models 1 and 2 using a chi-squared test, a test of the significance of the interaction term. We used an alpha level of 0.001 to Bonferroni adjust for the 44 items. To detect uniform DIF, the relative difference between the parameters associated with θ [β_1 from equations 2 and 3] was determined using the formula $|(\beta_{1(\text{equation 2})}-\beta_{1(\text{equation 3})})/\beta_{1(\text{equation 3})}|$. If the relative difference was greater than 10%, group membership interfered with the expected relationship between object naming ability and item responses (46) A 0.001 p-value criterion for uniform DIF can also be used with our -difwithpar- program.

We accounted for DIF by using items free of DIF as anchors; items found to have DIF had language-group specific item parameters estimated. The resulting estimates of object naming ability (θ) were thus generated from all of the items, but only DIF-free items had the same item parameters for the two languages. To account for spurious DIF, we used this DIF-free θ score as the ability level for DIF detection, and recalculated equations 1-3. We compared the items found to have DIF using the original $\hat{\theta}$ and using the modified $\hat{\theta}$. If the items found were the same, we concluded that our findings were not due to spurious DIF. If the items found with DIF were different, we used the most recent findings to generate new θ estimates, again using group-specific items when needed. These steps were repeated until the same items were found with DIF on successive runs. The final $\hat{q}$ values are IRT object naming scores that account for DIF. These steps are explained in more detail in Crane et al. (9)