Supplementary material 1

The experimental paradigm used in this study is based on Speech Reception Threshold (SRT) measurements. A given SRT is calculated over a number of trials (here eight) within an adaptive staircase (here using a 1-up/1-down rule) across a number of sentences (here ten) offering the same experimental condition. Because such paradigms capture variables affecting speech intelligibility, auditory masking, and source segregation in complex listening environments, researchers are generally not particularly interested in the participant's writing/spelling skills. Though there is arguably a fine line between literacy and speech comprehension skills, such paradigms tend to focus on the latter. An underlying assumption is that the participants themselves are in the best place to decide whether they correctly understood a given word (regardless of whether they know how to spell it). At first glance, this might seem like an untenable assumption considering that participants could intentionally cheat – especially when there is no oversight from any experimenter in an online setting – but such behaviors do not serve participants due to the adaptive nature of the task. The more a participant "cheats" (overestimating their score), the more adverse the TMR becomes, making the task more difficult for them. One consequence is that the staircase goes quickly towards difficult TMR levels, while the consequence of underestimating one's performance is to spend more trials at favorable TMRs. In both cases, the SRT extracted might deviate from the intended 50% point, essentially adding noise to the data. But it will not be biased in one direction or another unless there is systematic behavior from the participant.

In this Appendix, we addressed the possibility of participant unreliability by diligently screening all log files to verify the accuracy of participant self-scoring. In every single trial, we (manually) determined the number of keywords the participant should have written and compared it to their self-score, disregarding instances of close pronunciation similarity: homophones, e.g. tails/tales, slight differences of verb tense, e.g. run/ran or stop/stopped, and singular/plural forms. The French cohort made a total of 622 errors, representing 5.4% of trials, while the English cohort made a total of 637 errors, representing 6.9% of trials. These estimates are in close agreement to similar analyses conducted for in-laboratory experiments (5.7% in Deroche et al., 2017b), providing support for the decent quality of this online dataset. Therefore, the first confirmation finding is that self-scoring was largely accurate (green areas in Fig.S1, top panels). The results that follow are concerned with understanding the causes of these 5-7% errors.

Errors can be overestimations (red areas) or underestimations (blue areas). These errors are evidently spread differently, depending on the number of words reported (i.e. it is impossible to overestimate when reporting five words, or underestimate when reporting zero). In the bottom panels, the errors are further split by their size, keeping the same color code. While it is true that participants have a greater tendency to overestimate than underestimate their performance (about 80/20 for French cohort and 65/35 for the English cohort – see red/blue area ratios), the vast majority of errors is by +/- 1 word: 92% in the French cohort and 86% in the English cohort. Again, these estimates are in relatively close agreement to those reported for in-laboratory experiments (Deroche et al., 2017b) about 58/42 for the ratio of overestimation/underestimation, where 94% of errors were within +/-1 word. From this observation, we are inclined to suggest that these are "honest mistakes". Remember that the keywords are emphasized by being displayed in capitals. It is very easy indeed to miscount one of the lowercase words in the sentence as a false keyword or inversely thinking that one of the words typed in the response box

was too irrelevant to count. Evidence for cheating, or intentionally reporting erroneous scores, would be glimpsed by instances of +/- 3, 4, or 5 words errors, and there were only 7 and 15 cases of these larger errors in the whole dataset, representing 1% and 2% of errors (or 0.06% and 0.16% of trials) respectively in the English and French cohorts. In other words, there is very little evidence that participants intended to cheat; the very rare occasions where self-scoring was strikingly wrong could be operational errors (pressing a key by mistake).

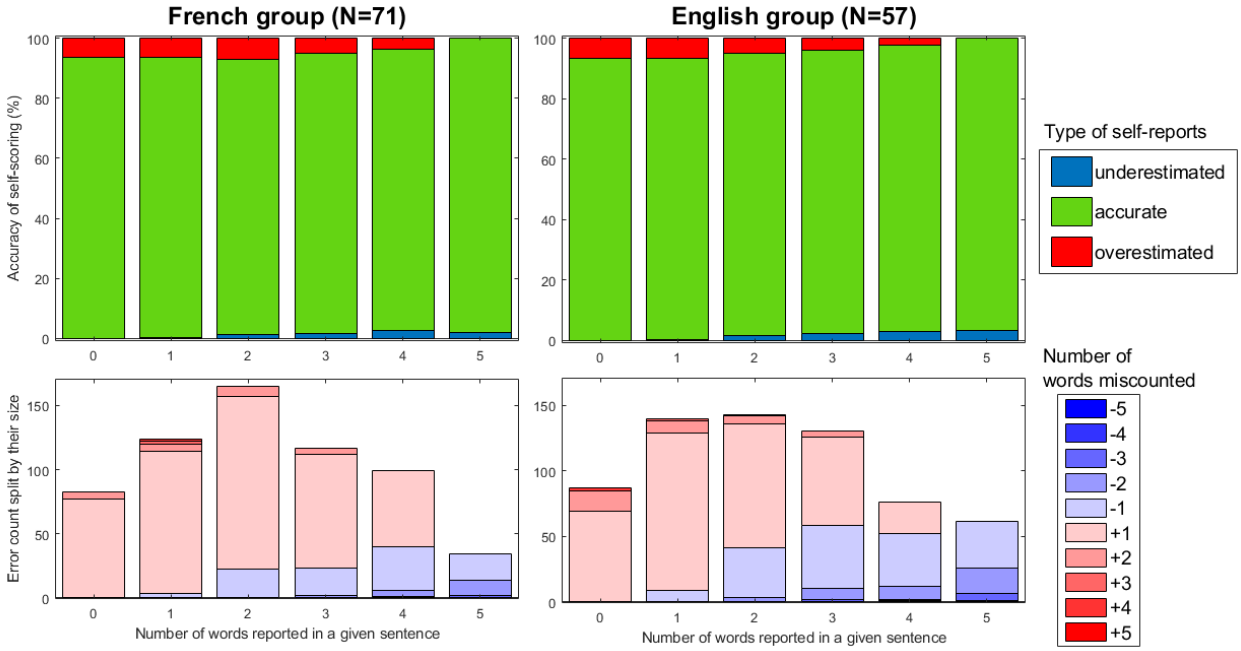

Figure S1. Graphical depiction of self-report accuracy (top panels) and severity of self-report errors (bottom panels) for both L1 French participants (left panels) and L1 English participants (right panels).

Another trait of these data that is worth mentioning (also reported in Deroche et al., 2017b) is that participants made fewer mistakes on the extremities of the scale (0 or 5) than in the middle of the scale (1-4), as illustrated by the height of the bars in the bottom panels (Fig.S1) regardless of color. This could genuinely reflect the added uncertainty that participants have when typing partial sentences only. The key point of this distribution is that *many errors made absolutely no difference to the data reported*, as they did not affect the adaptive staircase. While participants report a value between 0 and 5, this score is disregarded for the most part, since a value of 0, 1, or 2 is the same upward response (task becoming easier as target level increases by 2 dB), and a value of 3, 4, 5 is the same downward response (task becoming harder as target level decreases by 2 dB). In other words, all +2 errors when participants reported 0 word, as well as +1 or -1 errors when participants reported 1 or 4 words, and -2 errors when participants reported 5 words would be considered "correct" as far as the adaptive staircase is concerned. This is because SRT is calculated from the TMR at which the participant performs, not from the scores they report. The only errors that ultimately mattered are those that lead the staircase to a different path, i.e. errors that crossed the 2- versus 3-word boundary. These cases represented only 30% of errors in the French cohort and 30% in the English cohort (in close agreement to the estimate of 25% for

in-laboratory experiments in Deroche et al., 2017b). Put it differently, a given participant made around 3-4 mistakes on average over the whole study, which had the potential to affect the SRT measurement. Finally, if we consider that underestimation errors partly cancelled out overestimation errors, either within the same block or by averaging across participants, the potential for a systematic bias in our data becomes negligible.

This analysis opened new questions however – not exactly related to the present investigation – on the accuracy of self-scoring as a function of experimental condition. The following paragraph is no longer related to the reliability of the self-scoring method for SRT measurements but rather on the role of target and masker language for the participant's behavior in this self-scoring exercise, e.g. questions like "Would participants be more likely to miscount words that are in their L1 or in their L2? Would this behavior somehow change with different languages in the background?". To this aim, we extracted the averaged error per block and per subject (where underestimations and overestimations could cancel each other out), averaged it per condition, and used this DV in a mixed ANOVA with target and masker as within-subject factors and group (L1-French vs L1-English) as the between-subject factor. Suspecting that sex, nationality, and age (Deaux, 1979; Andrews, 1987; Pastorelli et al., 2001) could affect the tendency to underestimate or overestimate oneself, we added age and sex (nationality being already considered in Group) as covariates. Three participants were excluded: one because they did not report their sex, and the other two because they were outliers (strongly overestimating their scores while being among the oldest in our sample, resulting in dubious age effect). The final statistical results of these analyses are shown in Tables A1 and A2. No main effects or interactions were statistically significant in either analysis. This supports the idea that self-scoring was similar in bilinguals with French dominance and bilinguals with English dominance, and did not depend much on whether the target sentences were presented in their L1 or L2, nor whether the masker speech was L1, L2 or Lf. To conclude, the findings reported in this Appendix confirm that the self-scoring method was reliable and independent of the experimental condition.

Table A1: Error analysis of targets against noise maskers

Within-subject effects	
Target language	F(1, 121) = 0.322, p = .572
Target language × Age	F(1, 121) = 0.074, p = .786
Target language × Sex	F(1, 121) = 0.127, p = .722
Target language × Group	F(1, 121) = 1.487, p = .225
Between-subject effects	
Age	F(1, 121) = 0.080, p = .778
Sex	F(1, 121) = 0.010, p = .920
Group	F(1, 121) = 0.014, p = .904

Table A2: Error analysis of targets against speech maskers

Within-subject effects	
Target language	F(1, 121) = 3.238, p = .074
Target language × Age	F(1, 121) = 0.670, p = .415
Target language × Sex	F(1, 121) = 0.174, p = .677
Target language × Group	F(1, 121) = 0.086, p = .770
Masker language	F(2, 242) = 0.203, p = .817
Masker language × Age	F(2, 242) = 0.061, p = .941
Masker language × Sex	F(2, 242) = 2.953, p = .054
Masker language × Group	F(2, 242) = 1.183, p = .308
Target language × Masker language	F(2, 242) = 1.144, p = .320
Target language × Masker language × Age	F(2, 242) = 0.145, p = .865
Target language × Masker language × Sex	F(2, 242) = 0.361, p = .697
Target language × Masker language × Group	F(2, 242) = 1.969, p = .142
Between-subject effects	
Age	F(1, 121) = 0.191, p = .663
Sex	F(1, 121) = 1.816, p = .180
Group	F(1, 121) = 0.142, p = .707