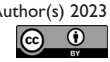
Article

LANGUAGE TEACHING RESEARCH

Language Teaching Research
2026, Vol. 30(4) 2085–2109



DOI: 10.1177/13621688231176067
journals.sagepub.com/home/ltr

'The wisdom of crowds': When teacher judgments outperform word-frequency as a predictor of students' vocabulary knowledge

Pablo Robles-García
University of Toronto, Mississauga, Canada

Jeffrey Stewart
Tokyo University of Science, Japan

Christopher Nicklin
Rikkyo University, Japan

Joseph P. Vitta
Kyushu University, Japan

Stuart McLean
Momoyama Gakuin University, Japan

Brandon Kramer
Kwansei Gakuin University, Japan

Abstract

This study investigated the effectiveness of word-frequency and teacher judgments in determining students' vocabulary knowledge and compared the predictive powers of both approaches when estimating vocabulary knowledge. Twenty-nine second language (L2) Spanish teachers were

Corresponding author:
Pablo Robles-García, Department of Language Studies, University of Toronto, Mississauga, 3359 Mississauga Rd, Maanjiwe Nendamowinan, 4th Floor, Mississauga, ON L5L 1C6, Canada
Email: p.roblesgarcia@utoronto.ca

asked to predict how likely their students would know words from a 216-word Yes/No test that measures knowledge of the first 3,000 words in Spanish. The accuracy of their responses was compared with the results of 1,075 L2 Spanish students who completed the same test. To examine if the results could generalize to other L2 settings, 394 L2 English students completed a 70-word Yes/No test that measures knowledge of the first 14,000 words in English, and 15 L2 English language instructors attempted to predict which words would or would not be recognized. Results showed that for both language contexts, (1) the median teacher rater could assess students' vocabulary knowledge with an accuracy roughly comparable to frequency, (2) the combination of teachers' judgments displayed a stronger relationship with students' performance on the vocabulary test than frequency, since the average of three or more teachers' ratings improved upon frequency when examined with 1,000 bootstrapped samples, and (3) using teacher judgments and frequency together did not substantially improve the prediction of students' vocabulary knowledge.

Keywords
bootstrapping, Spanish, teacher judgments, vocabulary knowledge, vocabulary selection, word-frequency

I Introduction

Some of the most significant challenges that language teachers face in second language (L2) vocabulary education are (1) deciding the most appropriate spoken and written texts that instructors should provide to their students to ensure adequate comprehension, (2) determining which words need to be replaced or glossed over in any given text, and (3) modifying the teachers' speech according to students' lexical proficiency. However, these decisions might prove a daunting task considering that the lexicon of most languages is composed of a network of tens of thousands of words (Schmitt et al., 2017). Furthermore, the fact that exposure to the L2 is often limited to the time students spend in the classroom makes vocabulary selection even more difficult to accomplish (Sánchez-Gutiérrez, Pérez Serrano, & Robles-García, 2019). Thus, language teachers might wonder what sources of information they can use to guide decisions about vocabulary selection. The purpose of this study is to examine how well teachers' own judgements are able to predict students' knowledge of words, and how their judgements compare to perhaps the most common predictor of word knowledge in L2 vocabulary pedagogy, word-frequency.

II Literature review

One of the most widely used principles for vocabulary selection in L2 teaching is the notion of lexical frequency (Horst, 2013; Schmitt & Schmitt, 2014). It is well known that a relatively small number of high-frequency words provide learners with the largest amount of lexical coverage in most spoken and written texts, and that such coverage decreases exponentially when words become less frequent (Davies & Hayward Davies, 2017; Nation, 2006). Previous scholarship confirms that the 3,000 most

frequent words[1] in a language offer approximately 95% of vocabulary coverage in most spoken and written texts (Davies, 2005; Nation, 2006; Schmitt & Schmitt, 2014), which provides learners with the vocabulary needed for adequate comprehension (Laufer & Ravenhorst-Kalovski, 2010; Van Zeeland & Schmitt, 2013). Furthermore, research has shown that there is a strong correlation between word-frequency (ranked by 1,000-word frequency bands) and students' vocabulary knowledge (Dóczi & Kormos, 2016; Milton, 2006a; Stæhr, 2008; Webb & Chang, 2012). Therefore, vocabulary researchers strongly encourage language teachers to follow a vocabulary selection criterion based on frequency (Horst, 2013; Stæhr, 2008), and to use corpus-based frequency as a means of determining the likelihood that learners will know a word (Milton, 2006b; Robles-García, 2022).

Despite the pedagogical recommendations proposed by the existing literature, students are far from being effectively exposed to the most frequent vocabulary in the L2 classroom. In most educational settings and university language courses around the globe, textbooks constitute one of the pillars of language-teaching curriculum design, especially when several sections from the same course are taught by different instructors simultaneously. Although the use of teacher-produced materials has experienced a significant increase in language teaching (McGrath, 2013), research has shown that textbooks strongly influence the pedagogical decisions made by language instructors (Cubillos, 2014; Neary-Sundquist, 2015).

When it comes to textbooks for English language teaching (ELT), publishers often claim to engage in robust content development practices supported by research (see review in Vitta, 2021). Cambridge University Press (2021), for instance, cites the English profile project (Green, 2008) when arguing for the suitability of their content. In fact, there have been empirical observations to support this claim in terms of frequency in particular (Crossley & McNamara, 2008) and grammar and vocabulary overall (Zarco-Tejada, 2019). Nevertheless, there have been instances of research highlighting problems with frequency in ELT textbooks. Nordlund (2016), for example, observed that ELT textbooks used in elementary Swedish contexts of English as a foreign language (EFL) overrepresented infrequent vocabulary, suggesting that the extent to which EFL published textbooks assist teachers with selecting and presenting appropriate (i.e. frequency-driven) vocabulary items is an open question. However, research on high-frequency words in L2 textbooks other than English has shown that vocabulary selection is not primarily based on lexical frequency, as evidenced by the underrepresentation of these words in most educational materials from different languages and proficiency levels alike (for German, see Lipinski, 2010; for Spanish, see Sánchez-Gutiérrez, Marcos Miguel, & Olsen, 2019).

In addition to this, language teachers often prefer to rely on their intuition rather than using empirical evidence to make pedagogical decisions for developing educational materials (Creighton, 2007; Kahneman & Frederick, 2005; Schildkamp & Ehren, 2013; Vanlommel et al., 2017). For instance, when it comes to instructors' beliefs about vocabulary selection, research with teachers of English (Dang & Webb, 2020) and Spanish (Sánchez-Gutiérrez et al., 2022) has shown that educators do not use corpus-based frequency to make decisions about the words they want to teach in the classroom. In Dang and Webb's (2020) study with 16 Vietnamese EFL teachers, the authors found that

corpus-based vocabulary lists were considered the least useful and influential sources of information for word selection. On the contrary, instructors reported that textbooks and their own intuitions were the main sources used for making decisions about what words to introduce in their lesson plans. Similar views on the use of corpus-based lists for vocabulary selection were also found in Sánchez-Gutiérrez et al.'s (2022) interviews of L2 Spanish teachers in the U.S. Although instructors from this study reflected on the potential benefits of using corpus-based vocabulary lists, they mentioned that these sources of information were not used for vocabulary selection due to lack of access or limited familiarity with available resources. They also reported basing many of their vocabulary selection decisions on textbooks and their own intuitions about the usefulness of a word. Instructors based word usefulness on their relevance when completing classroom activities, or on the applicability of those words to real-life situations. Finally, more than 50% of the instructors also mentioned that they would often try to use their own intuitions to find out which words were already known – and not known – by their students so they could base vocabulary selection on the words students did not know.

Results from these studies clearly indicate that teachers' decision-making is predominantly based on their own intuitions, rather than being guided by research-based recommendations. Therefore, considering the current pedagogical practices reported by language teachers, and given the strong correlation between word-frequency and students' vocabulary knowledge, it is necessary to examine how accurate teachers' intuitions are when predicting word-frequency.

Research on word-frequency judgments has mainly focused on native (Alderson, 2007; Carroll, 1966; Schmitt & Dunham, 1999; Shapiro, 1969) and non-native speakers (Arnaud, 1989; Kirsner et al., 1984; Schmitt & Dunham, 1999); while the accuracy of word-frequency judgments by L2 teachers has remained largely unexplored. Indeed, to the best of our knowledge, the study conducted by McCrostie (2007) is the only one that has focused on this subject. In an attempt to investigate the accuracy of teachers' word-frequency intuitions, the author analysed the accuracy of English as a foreign language (EFL) teachers' word-frequency judgments. To do so, 21 EFL teachers were asked to rank words from two different ranges of frequency: high-frequency words (words up to the first 2,000 most frequent words in English) and mid-frequency words (words among the 4,000–10,000 levels). Results showed that teachers' judgments were accurate when determining frequency from very high frequency words (i.e. first 1,000 words) and from the lowest mid-frequency levels (8,000–10,000), while showing serious difficulties when correctly ranking words from other mid-frequency levels.

As previously mentioned, teachers tend to rely on L2 textbooks and their own intuitions when deciding what words to teach in their language courses; and many believe that they have an intuitive understanding of which words their students know and do not know (McCrostie, 2007; Sánchez-Gutiérrez et al., 2022). However, recent work by McCrostie (2007) shows that (1) teachers only possess accurate judgments with words at both ends of the frequency spectrum (very-high-frequency and low-frequency words), and (2) L2 textbooks (mostly in L2s other than English) do not seem to foster a systematic approach to high-frequency words, since they introduce a large quantity of low-frequency words. Therefore, it seems reasonable to argue that teachers should consult and complement their own judgments with frequency-based wordlists for vocabulary

selection. Indeed, vocabulary selection will ultimately be based on what teachers believe about the usefulness of those words for their students. If frequency-based lists contain words that teachers do not consider to be useful for students, those words will not be introduced in the classroom. Consequently, several studies have advocated for the creation of word lists that consider frequency-based word data and teachers' intuitions about the usefulness of those words in real classroom settings (Dang et al., 2020; He & Godfroid, 2019; Stein, 2017).

In fact, these types of word lists have shown to provide better estimations of students' vocabulary knowledge than purely frequency-based vocabulary lists, as further evidenced in a recent study conducted by Dang et al. (2020). The authors examined the usefulness of Nation's (2012) British National Corpus / Corpus of Contemporary American English (BNC/COCA2000) and Brezina and Gablasova's (2015) New General Service List (New-GSL) using students' vocabulary knowledge and teacher perceptions of word usefulness. They decided to compare these two different word lists because both use word-frequency as their criterion for vocabulary selection. However, the BNC/COCA2000 also includes certain words that do not meet frequency criteria but have been considered appropriate for L2 learners (days of the week, months, numbers, common expressions such as *hello* and *goodbye*, names of countries, etc.). For the analysis, a total of 973 non-overlapping headwords were selected: 428 from the New-GSL and 545 from the BNC/COCA2000. Seventy-eight experienced EFL teachers were asked to indicate the usefulness of these 973 headwords in helping their students perform basic communicative situations in English using a five-point Likert scale. Likewise, 135 Vietnamese EFL learners completed a total of 15 Yes/No vocabulary tests that included the same 973 words from the teachers' survey. Results showed that the words selected from the BNC/COCA2000 were perceived as more useful by teachers and better known by students than those from the New-GSL. These findings suggest that the creation of effective vocabulary lists should be mainly frequency-based but complemented with teacher judgments. Indeed, the combination of these two constructs will allow for the development of vocabulary lists that follow frequency as the principled criterion but are organized based on teachers' judgments about their students' current vocabulary knowledge. Furthermore, these vocabulary lists might also introduce words that do not necessarily follow a frequency-based approach but are considered necessary for students to perform certain communicative tasks. This way, students will be exposed to useful real-world vocabulary that takes into account their particular vocabulary needs.

While the importance of developing these types of word lists has been addressed, one of the most important questions for vocabulary researchers and language practitioners alike is to know how effective (1) word-frequency and (2) teacher judgments are when predicting which words might be known (and not known) by L2 learners, and what source can be a better predictor of students' vocabulary knowledge.

For the last decades, word-frequency has been widely used as de facto methodology for identifying which words learners are more or less likely to know (Hashimoto, 2021; Meara, 2010; Schmitt et al., 2021), and has been the guiding principle for the development of most Vocabulary Size-and-Levels Tests to date (Hashimoto, 2021). However, in recent years, a number of researchers have thrown the accuracy of the vocabulary knowledge/frequency relationship into question. For instance, in a study conducted by Schmitt

et al. (2021) with L2 Spanish, German, and Chinese learners, the authors found a modest correlation between students' knowledge and frequency rankings (.23, .22, and .40 for Spanish, German, and Chinese, respectively). Likewise, similar findings were found in Hashimoto's (2021) study, in which the author found a moderate correlation ($r = .50$) between word difficulty (the probability that test-takers will know a word) and word-frequency. Nevertheless, it is worth noting that despite their obvious limitations, frequency-based lists are still one of the most powerful tools for predicting students' vocabulary knowledge.

Alongside textbooks and frequency lists, teacher judgments constitute another approach for selecting the vocabulary taught in language classrooms. However, teacher judgments have largely remained relatively unexplored. As previously mentioned, teachers are not active consumers of frequency-based wordlists for vocabulary selection; they tend to rely on their own judgments as a matter of expediency when selecting words to teach in the classroom (Dang et al., 2020; Sánchez-Gutiérrez et al., 2022), and many believe they possess an intuitive understanding of their students' word knowledge. Therefore, it seems reasonable to argue that teachers' perceptions about what students know, and do not know, will strongly influence the pedagogical decisions made in the classroom and should thus be analysed accordingly. Furthermore, given that most of the learning that takes place in the language classroom is conditioned by the activities selected by the instructors, it is possible that teacher intuition could be a competitive predictor of students' vocabulary knowledge, or even add explanatory power to predictions made by sources such as word-frequency lists. Indeed, while frequency may be a useful guideline for predicting vocabulary knowledge, teachers likely have a better understanding of words that students know but do not necessarily follow frequency-based parameters. For instance, *blackboard, marker, pencil*, or *folder* are low-frequency words that students are likely to know because they relate to classroom language content. On the contrary, words like *tendency, broadcast*, or *youth*, although highly frequent and extremely useful for students to become competent language users beyond classroom settings, might not be known by students because they fall outside of classroom-specific domains. Therefore, teacher judgments, if shown to be reliable measurements of students' vocabulary knowledge, could have pedagogical advantages over frequency or difficulty lists when determining what words among the frequent ones should be implemented in the classroom, since teacher judgements are easier and faster to collect than empirical data on students' vocabulary knowledge.

III Research questions

The goal of this study is to (1) measure the effectiveness of word-frequency and teacher judgments in determining students' vocabulary knowledge, and (2) compare the predictive powers of both methods. Concretely, this study responds to the following research questions:

- Research question 1: How do low, high, and median accuracy teacher judgments compare to frequency as a predictor of their students' vocabulary knowledge?

Accuracy of judgment ratings most likely will differ from teacher to teacher. Therefore, we will analyse how individual teachers' judgments of students' vocabulary knowledge perform on average and explore the upper and lower limits of predictive accuracy relative to the benchmark of predictions based on word-frequency.

- Research question 2: To what extent can teacher judgments be improved by the combination of multiple teachers?

Even in the event individual teacher judgments lag frequency as a predictor, it is possible that combining and averaging such judgments could improve accuracy by harnessing 'the wisdom of crowds' (Surowiecki, 2005). To do so, we will determine if accuracy can improve with larger numbers of teacher ratings and attempt to track how accuracy improves as more judges are added to the average, in order to determine when using more teacher judgments leads to diminishing returns.

- Research question 3: Can predictive accuracy be improved by using frequency and teacher judgments together?

As suggested above, teachers may have insights into which words their students do and do not know that are not available from frequency. Likewise, frequency may shed insight into this issue that is not available to most teachers. Therefore, it is possible that using the two variables in tandem could lead to greater predictive power than either variable alone.

These research questions will be answered with two different datasets with sufficient target words in each to detect effect sizes found in past word difficulty research (see Supplemental Material A). With dataset 1 we will analyse the accuracy of 29 L2 Spanish teachers in predicting how likely it is that their students will know words from 3K-LEx (Robles-García, 2020), a 216-word Yes/No vocabulary test that measures knowledge of the 3,000 most frequent words in Spanish. The teachers' judgments will be compared with the results of 1,476 L2 Spanish learners who were tested on the same 216 words from 3K-LEx. With dataset 2 we will analyse the accuracy of 15 English teachers when it comes to predicting the probability of 394 students knowing words chosen from the 14,000-word version (14k) of the Vocabulary Size Test (Nation & Beglar, 2007). This way, using two different datasets will help ensure the application of these findings not just between populations, but also between different L2 target vocabulary.

IV Dataset 1: L2 Spanish

1 Method

a Participants. In this study, dataset 1 participants for L2 Spanish came from two different groups: (1) 1,476 undergraduate Spanish L2 students from two U.S. public universities (1,105 female, 354 male, 17 other; ages between 18 and 58 years; $M = 20.04$, $SD = 2.35$), and (2) 29 language instructors who taught Spanish courses at these two institutions, accounting for more than 80% of the teacher population of both universities when data collection took place. From the student pool of 1,476 participants, 1,034

were first language (L1) English speakers, 71 were native speakers of Mandarin Chinese, and the remainder had other L1s, such as Arabic, Korean, Japanese, Russian, Urdu, Hindi, Cantonese, and Vietnamese, amongst others. In order to have a representative sample from the student population, the study was conducted with students from all the available Spanish courses at these two universities: first-year, second-year, and upper-level courses (Spanish literature, linguistics, and culture). To ensure higher levels of reliability in the students' responses, a 10% false alarm cutoff threshold was implemented in this study. Therefore, the results of participants who responded affirmatively to 11 (10.80%) or more of the 108 pseudowords included in the test were eliminated from the analysis. This procedure resulted in a final set of 1,075 participants. From these participants, 508 were taking first year courses, 328 were taking second year courses, and 239 were taking upper-level courses that focused on Spanish linguistics, literature, and culture. From the Spanish instructor pool (15 female, 14 male; ages between 23 and 52 years; $M = 32.4$, $SD = 6.96$), 14 were teaching first-year Spanish courses, 12 were teaching second-year courses, and three were teaching upper-level courses. A total of 23 instructors declared that Spanish was their native language, whereas six reported English as their first language.

b Instruments. Students completed the two available versions of 3K-LEx (Robles-García, 2020), a Yes/No lexical decision vocabulary levels test that measures knowledge of the 3,000 most frequent words in Spanish. Each version of the test contains a total of 162 items: 108 real words and 54 pseudowords. These words were randomly selected from Davies and Hayward Davies' (2017) Spanish lemmatized frequency dictionary, which is based on *Corpus del español* ('Spanish corpus'; Davies, 2002), a two-billion-word data based with a wide variety of spoken and written texts and a good balance of texts from Spain and Latin America.

Instructors, on the other hand, completed a survey developed to examine the teachers' perception of students' vocabulary knowledge of the 216 words from 3K-LEx. In this survey, instructors had to indicate how likely they thought their average student would know the words listed using a six-point Likert scale: 1 = Definitely doesn't know; 2 = Probably doesn't know; 3 = Might not know (slightly more likely the word is unknown); 4 = Might know (slightly more likely the word is known); 5 = Probably knows; 6 = Definitely knows. Although bootstrapping was employed in the main analysis, there was adequate psychometric justification for aggregating all judgments into a scale (see Supplemental Material B).

The observed reliability on the Spanish vocabulary instrument was strong (KR-20 = .99), and a principal components factor analysis, utilized as only one construct was theorized (see Stewart et al., 2022), suggested homogeneity via an observed 'elbow distribution' as the eigenvalue of the strongest factor (54.22) was approximately five times larger than the observed eigenvalue in the next strongest factor (10.74; see Osborne et al., 2008). Together, these observations implied the Spanish vocabulary instrument had good underlying psychometric properties.

c Procedure. Both groups of participants (instructors and students) completed their corresponding tests via Qualtrics (www.qualtrics.com), a web platform designed for data collection. Students took the test towards the end of the semester to ensure higher levels

of vocabulary knowledge. Teachers were also asked to complete the survey during the same period of time so their judgments about students' vocabulary knowledge would be as accurate as possible. Students completed the test during classroom time, and the test was administered both by one of the authors of this study and by instructors in their corresponding courses. In order to ensure a rigorous supervision of the test, instructors were given a short workshop and a set of detailed written instructions to familiarize with the nature of the task. Instructors were not asked to take the survey during classroom time, and they were allowed to complete it in their own time.

Before completing the test, all participants (students and instructors) signed a consent form and filled out a linguistic background questionnaire. Students were told that they were going to complete a Yes/No vocabulary test in which they had to read through a series of words and press *yes* if they knew the meaning of the word (if they were able to translate the word into their first language) and *no* if they did not know the meaning. Furthermore, they were encouraged to press *no* if they were not sure about the meaning of the word. Students were also told that the test would consist of two different lists of words, and that there would be a 3-minute break between the two lists. To reduce students' anxiety levels, they were notified that their responses would be de-identified and that their regular instructors would not have access to their results. No time restrictions were given to participants (with the exception of finishing the test during classroom time), and all students finished the task within 25 minutes.

Instructors were told that the purpose of this survey was to learn more about the accuracy of teacher judgments in predicting their students' current lexical knowledge. They were asked to think of an average student who was finishing their Spanish language course and were told that the purpose of this task was to decide whether they thought this average student would know words from a list of 216 words using the six-point Likert scale described previously. Likewise, teachers were not given any time restrictions, and they all completed the test within 26 minutes (except for one instructor who took about an hour to complete the test).

2 Results

a L2 Spanish students' vocabulary knowledge and frequency correlation. In order to respond to research question 1, the scatterplot illustrated in Figure 1 shows the initial correlation between students' vocabulary knowledge and log-transformed frequency. In the present study, vocabulary knowledge was operationalized as 'item facility', where a score of 1 indicates 100% of tested students reported knowing the word, and a score of 0 indicates none did. Frequency was log-transformed to account for the variable's Zipfian distribution (see Stewart et al., 2022). The resulting coefficient, $r = .67$ [.59, .74], demonstrated a strong relationship between vocabulary knowledge and frequency according to Plonsky and Oswald's (2014) guidelines.

b Individual teacher judgements and frequency correlations. Next, we examined how individual raters performed in their predictions by analysing the most and least accurate teachers. This was achieved by identifying which teacher's judgements displayed the strongest and weakest correlations with students' vocabulary knowledge. We also identified the

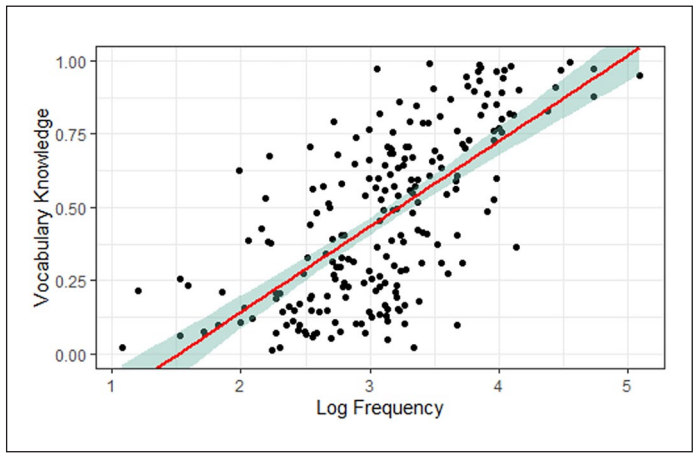


Figure 1. Scatterplot of Spanish word log frequency by vocabulary knowledge.

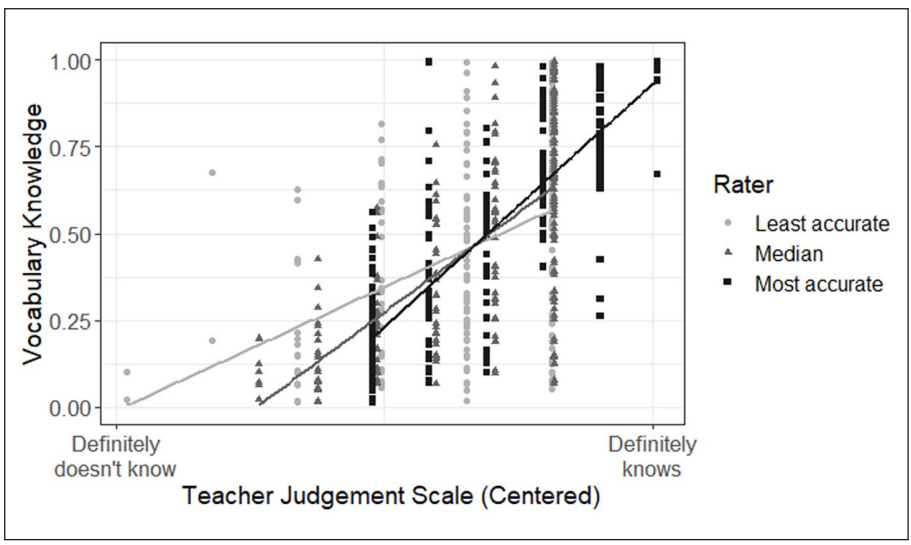


Figure 2. Scatterplots of lowest, highest, and median accuracy teacher judgements to vocabulary knowledge for Spanish data.

'average Joe', who was the teacher with the median correlation value. For each teacher, the judgment ratings for each word were converted to *z*-scores, which involved subtracting the mean score from each value and dividing the resulting number by the standard deviation (for the rationale, see Supplemental Material B).

Figure 2 indicates that the most accurate teacher (i.e. the teacher with the largest correlation coefficient, which is represented by the steepest slope) was more likely to assume

Table 1. Correlation coefficients between individual Spanish teachers' judgements and vocabulary knowledge.

Rater	Vocabulary knowledge	
	r	95% CI
Most accurate	.81	[.76, .85]
Median	.66	[.58, .73]
Least accurate	.41	[.29, .52]

Note. Log frequency by Vocabulary Knowledge $r = .67$ [.59, .74].

learners knew the words. As these predictions turned out to be accurate, the resulting correlation was very high, despite the restricted range of response values. In contrast, the least accurate teacher (i.e. shallowest slope) was more likely to assume learners did not know the words. Furthermore, the lack of overlap between the confidence intervals (CIs) presented in Table 1 indicate a significant difference between the most accurate rater and the 'average Joe', and between the 'average Joe' and the least accurate teacher.

c Aggregated teacher judgements and frequency correlations. As noted above, using multiple teacher judgements in unison may improve predictive power through 'the wisdom of crowds'. To determine if this is the case and thus answer research question 2, we standardized teacher responses to the Likert scale survey and combined and averaged judgment scores to determine if this process improved predictive power. We tested various numbers of combinations of raters, ranging from single raters to the maximum numbers of raters available in each dataset. A drawback of such an analysis is that the resulting correlations will likely differ depending on the precise raters used in each prediction, and the precise words examined in each case. To control for that, a bootstrapping method was employed. Bootstrapping is a robust statistical method in which numerous datasets are constructed from a single dataset to investigate how different combinations of variables affect the results of an analysis (i.e. Larson-Hall & Herrington, 2009; McLean et al., 2020). According to Plonsky et al. (2015), the simulated bootstrap samples offer stability and accuracy to estimates by providing a distribution of possible outcomes as opposed to a single-point measure.

For the Spanish data, 1,000 bootstrap datasets were sampled, with each bootstrap sample comprising 50 randomly sampled words and x number of teacher judgements. The teacher judgements were standardized, or centered to zero, and the mean rating for each word was calculated (for further explanation, see Supplemental Material B). In this way, we were able to simulate the 'wisdom of a crowd of teachers'. The students' vocabulary knowledge values for each word were correlated with the teachers' judgments and were also correlated with frequency. This process was repeated 1,000 times, and each time the Pearson's r correlation coefficients measuring the strength of the relationship between vocabulary knowledge and mean teacher judgment, and between vocabulary knowledge and frequency, were saved. This resulted in a set of 1,000 coefficients for each of the two correlational analyses. This process was repeated 14 times for groups of

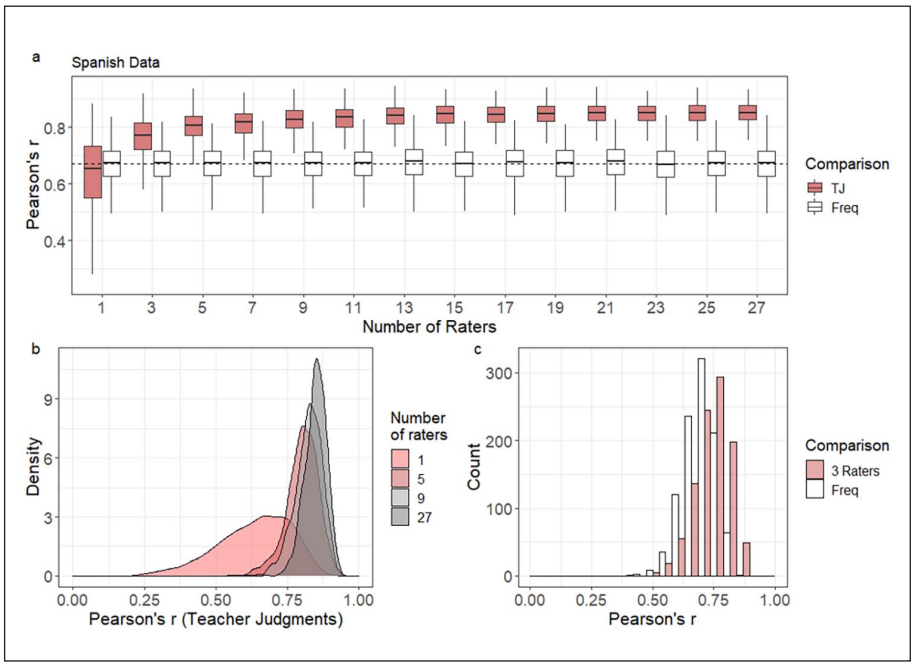


Figure 3. Plots of (a) vocabulary knowledge predicted using averages of teacher judgements and frequency with 1,000 bootstrapped samples, (b) bootstrap distributions of teacher judgment correlations by numbers of raters, and (c) bootstrap distributions of frequency correlations by three raters for Spanish data.

teachers in odd number intervals (i.e. 1, 3, 5, . . . 27). Bias-corrected and accelerated (BCa) confidence intervals, which account for skewness in a distribution, were calculated with the *coxed* R package (Kropko & Harden, 2020).

Results from this analysis (Figure 3 and Table 2) indicated that the combined scores of just three teachers were sufficient to achieve a stronger relationship with vocabulary knowledge than frequency in the bootstrapped simulations. For this dataset, a single rater's correlation with vocabulary knowledge produced an average of .64 [.32, .82], which was comparable to word-frequency, .67 [.50, .78]. This was substantially improved to .76 [.57, .87] when three raters were used and .80 [.67, .89] when five raters were used. The highest correlation, $r = .85$ [.75, .91], was achieved with 21 sets of judgements. However, a comparison between the coefficients and CIs for 21 teachers and 11 teachers, $r = .83$ [.73, .90], or even nine, $r = .82$ [.70, .90], reveals no practical difference, suggesting that as few as 10 sets of teacher judgements would be sufficient to create a variable that predicts word difficulty more precisely than frequency alone.

d Predicting students' vocabulary knowledge by combining teacher judgments and frequency. To explore how teacher judgments and frequency could complement each other in accounting for students' vocabulary knowledge (research question 3), a multiple regression model was constructed. The regression model and the preceding correlations'

Table 2. Descriptive statistics for 1,000 bootstrapped samples of teacher judgements and frequency correlated with vocabulary knowledge, by numbers of raters used (Spanish data).

Raters	Teacher judgments			Frequency		
	M	SD	95% CI	M	SD	95% CI
1	0.63	0.13	[.32, .82]	0.67	0.07	[.50, .78]
3	0.76	0.07	[.57, .87]	0.67	0.07	[.51, .78]
5	0.80	0.06	[.67, .89]	0.67	0.07	[.50, .77]
7	0.82	0.05	[.70, .90]	0.67	0.07	[.50, .78]
9	0.83	0.05	[.70, .90]	0.67	0.07	[.49, .78]
11	0.83	0.04	[.73, .90]	0.67	0.07	[.52, .78]
13	0.84	0.04	[.73, .90]	0.67	0.06	[.52, .78]
15	0.84	0.04	[.74, .90]	0.67	0.06	[.53, .77]
17	0.84	0.04	[.74, .91]	0.67	0.07	[.50, .78]
19	0.84	0.04	[.74, .91]	0.67	0.06	[.52, .77]
21	0.85	0.04	[.75, .91]	0.67	0.07	[.51, .78]
23	0.85	0.04	[.76, .91]	0.67	0.07	[.51, .77]
25	0.85	0.04	[.76, .91]	0.67	0.07	[.50, .77]
27	0.85	0.04	[.75, .91]	0.67	0.07	[.52, .78]

Table 3. L2 Spanish vocabulary knowledge predicted by frequency and teacher judgements.

Predictors	B	*SE*	β	*t*	*p*	B 95% CI		Pratt (%)	VIF
						Lower limits	Upper limits		
Frequency	0.12	0.02	0.28	4.70	< .001	0.07	0.17	12.76	1.79
Teacher judgments	0.28	0.02	0.72	15.54	< .001	0.25	0.31	61.42	1.79

Notes. Pratt (1987) Product Measure equates to β (standardized regression coefficient in the multivariate model) $\times$ *r* (correlation with the DV) for each predictor. Sum of Pratt values always equals R^2. VIF values < 2.50; multicollinearity was avoided (see Hashimoto & Egbert, 2019).

assumptions were met and are reported in Supplemental Material C. The model comprised 27 teacher judgments and frequency as predictors of vocabulary knowledge. The correlation between teacher judgments and L2 Spanish vocabulary knowledge was strong, $r = .85$ [.81, .88], indicating that teachers' judgments accounted for approximately 72% of the vocabulary knowledge variance, $r^2 = .72$. Additionally, the relationship between frequency and vocabulary knowledge was strong, $r = .67$ [.59, .74], although frequency accounted for less of the vocabulary knowledge variance than teacher judgements with approximately 45% explained, $r^2 = .45$. The lack of CI overlap when the entire dataset was considered implied that teacher judgments significantly outperformed frequency as a predictor of students' vocabulary knowledge (Cumming, 2012). In the final regression model reported in Table 3, $F[2, 213] = 305.98$, $p < .001$, the amount of explained vocabulary knowledge variance was approximately 74%, $R^2 = 74.18$. Analysis of Pratt (1987) values revealed that teacher judgments, 61.42%, contributed almost five times the amount of the explained variance in the model than frequency, 12.76%; see Table 3).

Interpreting these results in relation to research question 3, there is evidence suggesting the improvement of teacher judgments over frequency with their students. The observed R^2 of the regression model, 74% of the variance explained, is only 2% more than teacher judgments alone. There is evidence, however, of frequency having a minor role as it could significantly and independently contribute 12.76% of the explained variance alongside teacher judgements. Furthermore, considering that the bootstrapped coefficients indicated little difference between 9 and 27 teachers, these results suggest that approximately 10 sets of teacher judgements are sufficient to obtain a more accurate assessment of student knowledge than frequency.

V Dataset 2: L2 English

1 Method

In order to investigate the robustness of the applicability of the above results to other L2 settings, these analyses were re-conducted on a sample of Japanese university students studying English as a foreign language so that the population and target language were different.

a Participants. Similar to the previous study, participants came from two groups: (1) 394 undergraduate Japanese university students studying English from two private universities (179 female, 215 male; ages between 18 and 20 years); and (2) 15 language instructors who taught courses at these two institutions. The study was conducted with first-and second-year students taking oral communication classes. Using a 10% false alarm cutoff threshold, the results of participants who responded affirmatively to three (10%) or more of the 30 pseudowords were eliminated from the analysis, resulting in a final set of 313 participants.

The instructors were sampled from the teachers who taught in the departments where the classes were taught at the time of data collection. From the English instructor pool (two female, 13 male; ages between 29 and 70 years), seven were from school A and eight were from school B. One teacher reported Japanese as their first language, while the 14 other teachers reported English as their first language.

b Instruments. Students completed a 100-item Yes/No test which included 70 target words sampled from the first 14 levels of the vocabulary size test (VST; Nation & Beglar, 2007) and 30 pseudowords. The VST measures words sampled from the spoken subsection of the British National Corpus, ordered by frequency, and arranged into 1,000-word levels. The instrument's observed psychometric properties supported its use in the current study. The internal reliability was acceptable (KR-20 = .83) and there was evidence of homogeneity via an elbow distribution (as with the Spanish instrument) of the PCA results where the first factor's eigenvalue (= 6.33) was approximately two and a half times the value of the second (= 2.64).

Instructors completed a survey developed to examine the teachers' perception of students' vocabulary knowledge of the 70 words from the Yes/No test. In this survey, instructors had to indicate the likelihood that their average student would know the words listed using a six-point Likert scale: 1 = Definitely doesn't know; 2 = Probably doesn't

know; 3 = Might not know (slightly more likely the word is unknown); 4 = Might know (slightly more likely the word is known); 5 = Probably knows; 6 = Definitely knows. As with the L2 Spanish judgments, there was adequate justification for all teacher judgements representing a single construct (see Supplemental Material B).

c Procedure. Both groups of participants (i.e. the instructors and students) completed their corresponding tests online using SurveyMonkey. Students completed the test during classroom time, and the test was administered by two of the authors in this study. Instructors were asked to take the survey separately by email.

Before completing the test, all participants, including students and instructors, signed a consent form and were given the same instructions as the Spanish learners. They were notified that their responses would be de-identified and that their participation would not influence their grades. No time restrictions were given to participants, with the exception of finishing the test during classroom time, and all students finished the task within 15 minutes.

Instructors were told that the purpose of this survey was to learn more about the accuracy of teacher judgments in predicting their students' current lexical knowledge. They were asked to think of an average student in their communication classes and needed to decide whether they thought this average student would know words from a list of 70 words using the six-point Likert scale presented above. Teachers were not given any time restrictions, and they all completed the test within 10 minutes.

2 Results

a L2 English vocabulary knowledge and frequency correlation. The analyses conducted for the Spanish data were re-run for the English data. Figure 4 shows the initial correlation between vocabulary knowledge and log-transformed frequency, $r = .79$ [.68, .86], demonstrating a strong relationship between the variables according to Plonsky and Oswald's (2014) guidelines.

b Individual teacher judgements and frequency correlations. Figure 5, which plots the least accurate, 'average Joe', and most accurate raters, shows that the distinction is less clear than for the Spanish data (see Figure 2). The raters' judgements were more synchronized with no significant difference between the most and least accurate raters, as attested by the CIs in Table 4. This was most likely attributable to the large spread of vocabulary levels that the target words in the dataset were sourced from. It is easier to say that learners are unlikely to know words from less frequent vocabulary levels, such as *gimmick, ubiquitous*, or *atoll*, that constituted the bulk of the sample. Table 4 also suggests that the teacher judgments procured from the items in the English vocabulary test were more closely aligned with frequency than for the Spanish test results, which was also attributable to the large spread of vocabulary levels.

c Aggregated teacher judgements and frequency correlations. As with the Spanish dataset, 1,000 bootstrap datasets were sampled for the English data, with each bootstrap sample comprising 50 randomly sampled words and x number of teacher judgements. This analysis was done to investigate if the results found in Spanish were consistent, and thus applicable to another L2 context, in this case, L2 English teachers in Japan.

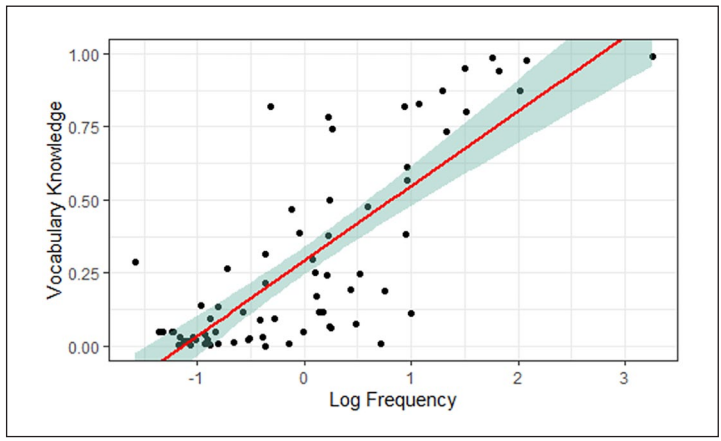


Figure 4. Scatterplot of English word log frequency by vocabulary knowledge.

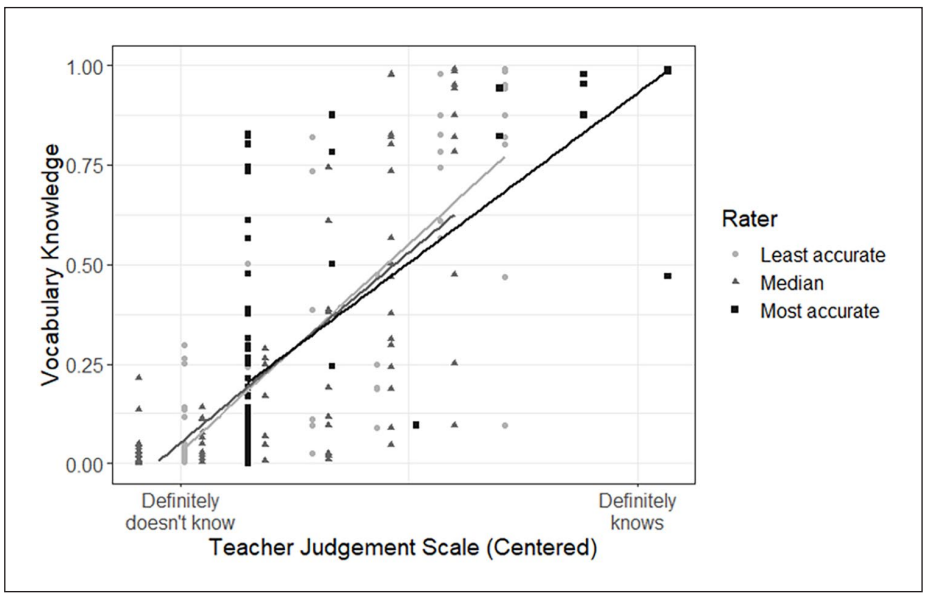


Figure 5. Scatterplots of lowest, highest, and median accuracy teacher judgements to vocabulary knowledge for English data.

Results from this analysis (Figure 6 and Table 5) also indicated that the combination of only five teachers was enough to improve the correlation between students' vocabulary knowledge and frequency. For the English set, a single rater's correlation with vocabulary knowledge, $r = .73$ [.60, .83], was weaker than the correlation found between students' vocabulary knowledge and frequency, $r = .78$ [.70, .84]. However, this

Table 4. Correlation coefficients between individual English teacher judgements and vocabulary knowledge.

Rater	Vocabulary knowledge	
	r	95% CI
Most accurate	.80	[.70, .87]
Median	.73	[.60, .83]
Least accurate	.65	[.49, .77]

Note. Log Frequency by Vocabulary Knowledge $r = .79$ [.68, .86].

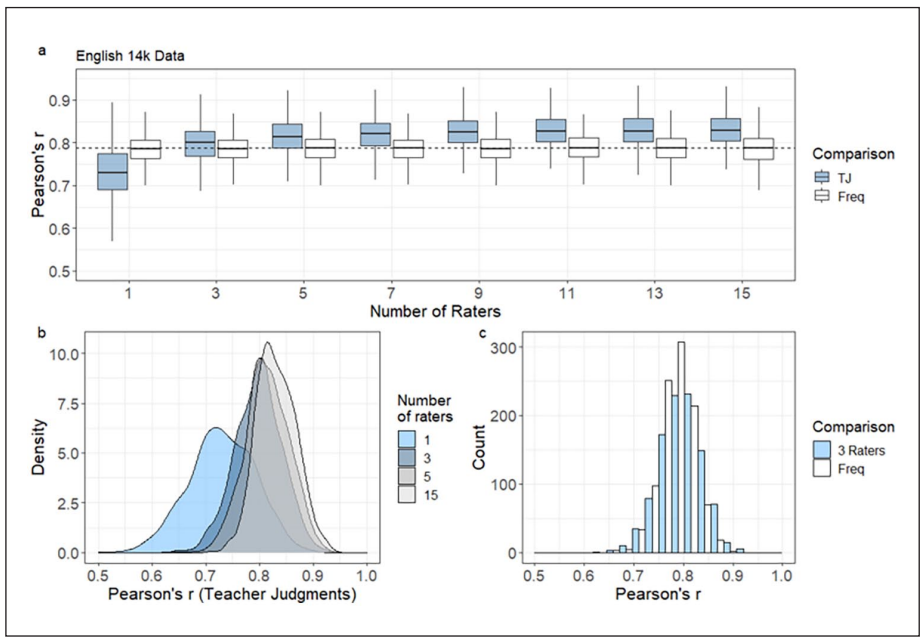


Figure 6. Plots of (a) vocabulary knowledge predicted using averages of teacher judgements and frequency with 1,000 bootstrapped samples, (b) bootstrap distributions of teacher judgment correlations by numbers of raters, and (c) bootstrap distributions of frequency correlations by three raters for English 14,000-word (14k) data.

correlation substantially improved to $r = .80$ [.72, .88] when three raters were introduced in the analysis and to $r = .82$ [.73, .89] when five raters were introduced. Nevertheless, with this vocabulary test the English raters achieved the strongest relationship with nine teachers $r = .83$ [.76, .91], although the difference between five and nine raters displayed no practical difference. In contrast, the Spanish raters required up to 21 teachers to achieve the strongest relationship, although no meaningful difference was observed whether 11 or 21 sets of Spanish rater judgments were procured.

Table 5. Descriptive statistics for 1,000 bootstrapped samples of teacher judgements and frequency correlated with vocabulary knowledge, by numbers of raters used (English data).

Raters	Teacher judgments			Frequency		
	M	SD	95% CI	M	SD	95% CI
1	0.73	0.06	[.60, .83]	0.78	0.03	[.70, .84]
3	0.80	0.04	[.72, .88]	0.79	0.03	[.71, .84]
5	0.82	0.04	[.73, .89]	0.78	0.03	[.71, .85]
7	0.82	0.04	[.74, .90]	0.79	0.03	[.71, .85]
9	0.83	0.04	[.76, .90]	0.79	0.03	[.71, .85]
11	0.83	0.04	[.76, .91]	0.79	0.03	[.71, .85]
13	0.83	0.04	[.77, .91]	0.79	0.03	[.71, .85]
15	0.83	0.04	[.76, .91]	0.79	0.04	[.71, .85]

Table 6. L2 English vocabulary knowledge predicted by frequency and teacher judgements.

Predictors	B	*SE*	β	*t*	*p*	B 95% CI		Pratt (%)	VIF
						Lower limits	Upper limits		
Frequency	0.13	0.03	0.36	3.75	$< .001$	0.06	0.19	28.40	2.44
Teacher judgments	0.21	0.04	0.56	5.77	$< .001$	0.14	0.28	46.15	2.44

Notes. VIF values < 2.50. Multicollinearity was avoided.

d Predicting students' vocabulary knowledge by combining teacher judgments and frequency. The results obtained with the VST target word set were similar to those observed with the Spanish set when assessing how teacher judgments and frequency performed in tandem (see Table 6). Frequency displayed a strong relationship with L2 English vocabulary knowledge, $r = .79$ [.68, .86], with approximately 62% of the variance explained, $r^2 = .62$. The relationship between teacher judgments and students' vocabulary knowledge was slightly, but not significantly stronger, $r = .83$ [.74, .89], and accounted for approximately 69% of the explained variance, $r^2 = .69$. In the regression model, $F(2, 67) = 98.18$, $p < .001$, approximately 75% of the variance was explained, $R^2 = 74.55\%$. As with the Spanish data multiple regression model, the Pratt values observed in the regression predicting L2 English vocabulary knowledge pointed to teacher judgments, 46.15%, having more predictive power within the model than frequency, 28.40%. The regression model and correlations' assumptions were met and are reported in Supplemental Material C.

As with the Spanish data, these results suggest that teacher judgments could be superior to frequency in predicting their students' vocabulary knowledge. In relation to research question 3, the data suggests that teacher judgements predict students' vocabulary knowledge sufficiently alone (compared to frequency) but there is also evidence for frequency having a role. Comparing the observed R^2 values of the regression model and teacher judgments, the former only accounts for 5% more of the variance. In the model

the teacher judgments measurement does outweigh frequency but to a lesser extent. This could be attributed to the greater frequency range in the VST target set compared to the L2 Spanish target words, whereby the very low frequency words in the 10,000 band and over were mostly unknown by the participants, and both frequency and teacher judgements predicted this resulting in less difference between them than observed in the Spanish data.

VI Discussion

The goal of this study was to explore word-frequency and teacher judgments as predictors of students' vocabulary knowledge, with the final aim of comparing the predictive powers of both methods when estimating vocabulary knowledge.

Concerning research question 1, results showed that there were great differences between the most, the median (i.e. 'average Joe'), and the least accurate teacher judgments when predicting vocabulary knowledge. Indeed, the correlation between the most accurate teacher and vocabulary knowledge was .81 and .80 for the Spanish and English data, respectively, surpassing the accuracy estimates of word-frequency. On the other hand, the correlation between vocabulary knowledge and the median teacher varied significantly between datasets. On the Spanish dataset, the median teacher's correlation was .66, and thus comparable to word-frequency, $r = .67$. However, the median teacher's correlation on the English data was .73, which was marginally weaker than students' vocabulary knowledge and frequency, $r = .79$. Finally, correlations between the least accurate teacher and vocabulary knowledge were also noticeably lower than word-frequency, with .41 and .65 for Spanish and English, respectively. These results suggest that single-teacher judgments should be interpreted with caution due to the large variation in correlational strength, and that teachers should be aware of the importance of using word-frequency lists when predicting students' vocabulary knowledge. Indeed, although some teachers might be better at predicting vocabulary knowledge than frequency itself, the large variation among them, and the fact that the average and least accurate teachers were at times less effective predictors of students' vocabulary knowledge than word-frequency suggests that frequency lists should be prioritized when predicting vocabulary knowledge, especially when predictions have to be made by teachers on their own.

However, how does the situation change when teachers work in tandem to predict their students' word knowledge? This issue was addressed in research question 2, and the results showed that the combined judgments of only three teachers were necessary to improve upon frequency in predicting vocabulary knowledge on both datasets. Indeed, the correlations between teacher judgements and vocabulary knowledge increased to .76 and .80 for Spanish and English, respectively, when the judgements of three teachers were combined, surpassing the accuracy estimates of word-frequency, .67 and .79. In addition, stronger correlations between teacher judgments and vocabulary knowledge were found as more instructors were introduced. For the Spanish data, a total of 21 teachers were needed to achieve the strongest relationship, $r = .85$, whereas the English data only needed nine teachers to achieve the highest relationship, $r = .83$. Additionally, the multiple regressions for both datasets addressing research question 3, where frequency

and teacher judgments worked in tandem, presented effect sizes that were only marginally better than the predictive power of teacher judgments alone.

Results from this study indicate that aggregated teacher judgments offer language teachers and program managers a simple and intuitive yet strong way to decide if students are likely to know target vocabulary. In fact, the combination of teacher judgments, whether used alone or in tandem with corpus-derived frequency values, offer vocabulary knowledge estimations that go above and beyond frequency, with at least two reasons to support this. First, it is worth mentioning that we specifically canvassed teacher judgments from teachers of the courses from which the data was collected, so the teachers knew the proficiency and performance capabilities of these students. As a result, teachers were able to know their students better than corpus-derived lists.

Second, corpus-derived frequency lists, although useful tools for predicting students' vocabulary knowledge, might not portray students' knowledge of words that are frequent in language teaching contexts but do not follow frequency-based parameters (Brysbaert et al., 2021). For instance, the word *folder* ranks 5,670 on the COCA wordlist. Based on frequency, *folder* is less likely to be known than *management* that ranks 1,040 in the same wordlist. However, L2 learners are more likely to know *folder* than *management* because *folder* is strongly related to classroom language content. Teachers, contrary to frequency lists, are aware of such phenomena and will thus often be better at predicting vocabulary knowledge than frequency. Consequently, this study supports recent research that has pointed out the need to consider not only frequency but teachers' intuitions about word usefulness and knowledge when developing vocabulary lists for teaching purposes (Dang et al., 2020; He & Godfroid, 2019; Stein, 2017).

This study's findings support the incorporation of teacher intuitions into the development of classroom vocabulary lists. As previously mentioned, frequency-based vocabulary lists offer instructors the tools they need to expose their students to vocabulary that will be essential for understanding a wide variety of L2 spoken and written texts with the final aim of helping them become competent language users. However, it has been shown that for the purpose of predicting students' vocabulary knowledge, teacher judgments can outperform corpus-based frequency lists. Therefore, for teaching purposes, a case can be made that the creation of vocabulary-lists should involve both vocabulary selection and teacher judgments about the knowledge of those words among their students. Indeed, teacher intuitions can be an important tool for fine-tuning frequency-based lists, since teachers will be able to effectively disregard words that students are likely to know so that the classroom word-lists can focus on words that students most likely do not know.

This study has provided robust evidence about the effectiveness of teacher judgments as useful predictors of students' vocabulary knowledge, however, there is one main limitation that should be addressed and taken into consideration. Although Yes/No vocabulary tests have been used in a wide variety of studies to date (Dang et al., 2020; Hashimoto, 2021; Hashimoto & Egbert, 2019; Nakata et al., 2020; Robles-García, 2022; Vitta et al., 2023), recent research has shown that other modalities such as meaning-recall vocabulary tests appear to be more accurate tools for measuring students' actual vocabulary knowledge in relation to predicting language skills such as reading (McLean et al., 2020; Zhang & Zhang, 2020). Indeed, Yes/No lexical decision tests have been criticized for overestimating vocabulary knowledge (Pellicer-Sánchez & Schmitt, 2012; Stewart &

White, 2011), since learners only need to decide whether they recognize the meaning of a given word without demonstrating actual knowledge of its meaning. Although the Yes/No test used in the present study displayed good reliability and also made these results consistent and comparable to previous studies, there would be value in further examining this issue using meaning-recall tests as the measure of learner vocabulary knowledge.

VII Conclusions

This study aimed to explore the effectiveness of word-frequency and teacher judgments in determining students' vocabulary knowledge while comparing the predictive powers of both methods. To do so, L2 Spanish and English students completed two different vocabulary tests. Those responses were then compared with the judgments of students' word knowledge reported by L2 Spanish and English instructors teaching those courses.

Results showed that for both the L2 Spanish and English datasets, (1) there was a strong relationship between vocabulary knowledge and frequency, (2) the accuracy of teacher judgments when predicting vocabulary knowledge varied widely among instructors, (3) the combination of teacher judgments correlated to actual students' vocabulary knowledge at higher levels than frequency, since the average of three or more teachers' standardized judgments exceeded word-frequency as a predictor of vocabulary knowledge, and (4) the combination of teacher judgments and frequency did not substantively improve the predictive power of students' vocabulary knowledge.

In conclusion, this study makes an important contribution to the field of vocabulary teaching and learning, since it offers insight into the effectiveness of teacher judgments, which has been a neglected area of research despite being a commonly used strategy for vocabulary selection in the L2 classroom. Furthermore, the replication of these results across different populations, languages, and ranges of word-frequencies suggests that the process demonstrated here could be applied to different L2 contexts. Finally, this study reinforces the need to refine frequency-vocabulary lists taking teachers' judgments of students' vocabulary knowledge as a major source of information, and urges the research community to give teachers' judgments the importance they deserve when creating such lists for teaching purposes.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Pablo Robles-García https://orcid.org/0000-0003-4780-8552
Jeffrey Stewart https://orcid.org/0000-0002-3350-3160
Christopher Nicklin https://orcid.org/0000-0002-8945-0678
Joseph P. Vitta https://orcid.org/0000-0002-5711-969X
Stuart McLean https://orcid.org/0000-0002-7035-378X
Brandon Kramer https://orcid.org/0000-0003-3910-0810

Supplemental material

Supplemental material for this article is available online.

Note

1. Although lexical coverage has been defined using different word units – lemmas for Spanish in Davies (2005) and word-families for English in Nation (2006) – research shows that readers need 3,000–4,000 words to ensure 95% of vocabulary coverage in both languages regardless of the word counting unit employed. This is due to the fact that most of the words within a family are not frequent enough to be considered high-frequency words. However, these word units will not provide similar lexical coverage when moving beyond high-frequency words (Robles-García et al., 2022).

References

Alderson, J.C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*, 383–409.

Arnaud, P.J.L. (1989). Estimations subjectives des fréquences des mots [Subjective estimates of word frequencies]. *Cahiers de Lexicologie*, *54*, 69–81.

Brezina, V., & Gablasova, D. (2015). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*, 1–22.

Brysbaert, M., Keuleers, E., & Mandera, P. (2021). Which words do English non-native speakers know? New supranational levels based on yes/no decision. *Second Language Research*, *37*, 207–231.

Cambridge University Press. (2021). *2021 ELT Cambridge University Press international catalogue*. Cambridge University Press.

Carroll, J.B. (1966). Quelques mesures subjectives en psycholinguistique: Fréquence des mots, significativité et qualité de traduction [Some subjective measures in psycholinguistics: Word frequency, significance, and translation quality]. *Bulletin de Psychologie*, *19*, 580–592.

Creighton, T.B. (2007). *School and data: The educator's guide for using data to improve decision making*. 2nd edition. Corwin Press.

Crossley, S.A., & McNamara, D.S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of Crossley, Louwerse, McCarthy, and McNamara (2007). *Language Teaching*, *41*, 409–429.

Cubillos, J.H. (2014). Spanish textbooks in the US: Enduring traditions and emerging trends. *Journal of Spanish Language Teaching*, *1*, 205–225.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.

Dang, T.N.Y., & Webb, S. (2020). Vocabulary and good language teachers. In Griffiths, C., & Z. Tajeddin (Eds.), *Lessons from good language teachers* (pp. 203–218). Cambridge University Press.

Dang, T.N.Y., Webb, S., & Coxhead, A. (2020). Evaluating lists of high-frequency words: Teachers' and learners' perspectives. *Language Teaching Research*, *26*, 617–641.

Davies, M. (2002). *El corpus del español [Spanish corpus]*. Available at: http://www.corpusdelespanol.org (accessed May 2023).

Davies, M. (2005). Vocabulary range and text coverage: Insights from the forthcoming Routledge frequency dictionary of Spanish. In Eddington, D. (Ed.), *Selected proceedings of the 7th Hispanic Linguistics Symposium* (pp. 106–115). Cascadilla Proceedings Project.

Davies, M., & Hayward Davies, K. (2017). *A frequency dictionary of Spanish: Core vocabulary for learners*. 2nd edition. Routledge.

Dóczi, B., & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. Oxford University Press.

Green, A. (2008). English profile: Functional progression in materials for ELT. *Research Notes*, *33*, 19–25. Cambridge ESOL.

Hashimoto, B.J. (2021). Is frequency enough? The frequency model in vocabulary size testing. *Language Assessment Quarterly*, *18*, 171–187.

Hashimoto, B.J., & Egbert, J. (2019). More than frequency? Exploring predictors of word difficulty for second language learners. *Language Learning*, *69*, 839–872.

He, X., & Godfroid, A. (2019). Choosing words to teach: A novel method for vocabulary selection and its practical application. *TESOL Quarterly*, *53*, 348–371.

Horst, M. (2013). Mainstreaming second language vocabulary acquisition. *The Canadian Journal of Applied Linguistics*, *16*, 171–188. Available at: https://journals.lib.unb.ca/index.php/CJAL/article/view/21299 (accessed May 2023).

Kahneman, D., & Frederick, S. (2005). A model of heuristic judgment. In Holyoak, K.J., & R.G. Morrison (Eds.), *The Cambridge handbook of thinking and reasoning* (pp. 267–293). Cambridge University Press.

Kirsner, K., Smith, M.C., Lockhart, M.L., & Jain, M. (1984). The bilingual lexicon: Specific units in an integrated network. *Journal of Verbal Learning and Verbal Behavior*, *23*, 519–539.

Kropko, J., & Harden, J.R. (2020). *coxed*: An R package for computing duration-based quantities from the Cox proportional hazards model. *The R Journal*, *11*, 38–45.

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied linguistics. *Applied Linguistics*, *31*, 368–390.

Laufer, B., & Ravenhorst-Kalovski, G.C. (2010). Lexical threshold revisited: Lexical coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15–30. Available at: https://nflrc.hawaii.edu/rfl/item/206 (accessed May 2023).

Lipinski, S. (2010). A frequency analysis of vocabulary in three first-year textbooks of German. *Die Unterrichtspraxis / Teaching German*, *43*, 167–174.

McCrostie, J. (2007). Investigating the accuracy of teachers' word frequency intuitions. *RELC Journal*, *38*, 53–66.

McGrath, I. (2013). *Teaching materials and the roles of EFL/ESL teachers: Practice and theory*. A&C Black.

McLean, S., Stewart, J., & Batty, A.O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing*, *37*, 389–411.

Meara, P. (2010). EFL vocabulary tests. Available at: http://www.lognostics.co.uk/vlibrary/meara1992z.pdf (accessed May 2023).

Milton, J. (2006a). Language lite? Learning French vocabulary in school. *Journal of French Language Studies*, *16*, 187–205.

Milton, J. (2006b). X-Lex: The Swansea vocabulary levels test. In Coombe, C., Davidson, P., & D. Lloyd (Eds.), *Proceedings of the 7th and 8th current trends in English language testing (CTELT) conference: Volume 4* (pp. 29–39). TESOL Arabia, UAE.

Nakata, T., Tamura, Y., & Aubrey, S. (2020). Examining the validity of the LexTALE test for Japanese college students. *The Journal of Asia TEFL*, *17*, 335–348.

Nation, I.S.P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82.

Nation, I.S.P. (2012). The BNC/COCA word family lists. Available at: http://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed May 2023).

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*, 9–13. Available at: https://jalt-publications.org/tlt/issues/2007-07_31.7 (accessed May 2023).

Neary-Sundquist, C.A. (2015). Aspects of vocabulary knowledge in German textbooks. *Foreign Language Annals*, *48*, 68–81.

Nordlund, M. (2016). EFL textbooks for young learners: A comparative analysis of vocabulary. *Education Inquiry*, *7*, 47–68.

Osborne, J., Costello, A., & Kellow, J. (2008). Best practices in exploratory factor analysis. In Osborne, J. (Ed.), *Best practices in quantitative methods* (pp. 86–99). Sage.

Pellicer-Sánchez, A., & Schmitt, N. (2012). Scoring Yes–No vocabulary tests: Reaction time vs. nonword approaches. *Language Testing*, *29*, 489–509.

Plonsky, L., Egbert, J., & LaFlair, G.T. (2015). Bootstrapping in applied linguistics: Assessing its potential using shared data. *Applied Linguistics*, *36*, 591–610.

Plonsky, L., & Oswald, F.L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Pratt, J. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. Unpublished Paper presented at the Proceedings of the Second International Tampere Conference in Statistics, Tampere, Finland: University of Tampere, pp. 245–260.

Robles-García, P. (2020). 3K-LEx: Desarrollo y validación de una prueba de amplitud léxica en español [3K-LEx: Development and validation of a lexical breadth test in Spanish]. *Journal of Spanish Language Teaching*, *7*, 64–76.

Robles-García, P. (2022). Receptive vocabulary knowledge in L2 learners of Spanish: The role of high frequency words. *Foreign Language Annals*, *55*, 963–984.

Robles-García, P., Wallace, G.H., & Sánchez-Gutiérrez, C. (2022). 3K-LEx-MC: The creation and validation of a multiple-choice vocabulary placement test for Spanish language learners. *ITL - International Journal of Applied Linguistics*. Epub ahead of print 25 October 2022. DOI: 10.1075/itl.22008.rob

Sánchez-Gutiérrez, C., Marcos Miguel, N., & Olsen, M. (2019). Words and textbooks: An analysis of vocabulary coverage and lexical characteristics in L2 Spanish textbooks. In Ecke, P., & S. Rott (Eds.), *Understanding vocabulary learning and teaching: Implications for language program development* (pp. 78–98). Cengage.

Sánchez-Gutiérrez, C., Pérez Serrano, M., & Robles-García, P. (2019). The effects of word frequency and typographical enhancement on incidental vocabulary learning in reading. *Journal of Spanish Language Teaching*, *6*, 14–31.

Sánchez-Gutiérrez, C., Robles-García, P., & Pérez Serrano, M. (2022). L2 Spanish vocabulary in US universities: Instructors' beliefs and reported practices. *Language Teaching Research*. Epub ahead of print 31 January 2022. DOI: 10.1177/13621688221074443

Schildkamp, K., & Ehren, M. (2013). From intuition to data-based decision making in Dutch secondary schools? In Schildkamp, K., Kuin Lai, M., & L. Earl (Eds.), *Data-based decision making in education* (pp. 193–207). Springer.

Schmitt, N., Cobb, T., Horst, M., & Schmitt, D. (2017). How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation, (2006), and Cobb (2007). *Language Teaching*, *50*, 212–226.

Schmitt, N., & Dunham, B. (1999). Exploring native and non-native intuitions of word frequency. *Second Language Research*, *15*, 389–411. Available at: https://www.jstor.org/stable/43111242 (accessed May 2023).

Schmitt, N., Dunn, K., O'Sullivan, B., Anthony, L., & Kremmel, B. (2021). Introducing knowledge-based vocabulary lists (KVL). *TESOL Journal*, *12*, 1–14.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*, 484–503.

Shapiro, B.J. (1969). The subjective estimation of relative word frequency. *Journal of Verbal Learning and Verbal Behaviour*, *8*, 248–251.

Stæhr, L.S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152.

Stein, G. (2017). Some thoughts on the issue of core vocabularies: A response to Vaclav Brezina and Dana Gablasova: 'Is there a core general vocabulary?' Introducing the new general service list. *Applied Linguistics*, *38*, 759–763.

Stewart, J., Vitta, J.P., Nicklin, C., McLean, S., Pinchbeck, G.G., & Kramer, B. (2022). The relationship between word difficulty and frequency: A response to Hashimoto (2021). *Language Assessment Quarterly*, *19*, 90–101.

Stewart, J., & White, D.A. (2011). Estimating guessing effects on the vocabulary levels test for differing degrees of word knowledge. *TESOL Quarterly*, *45*, 370–380.

Surowiecki, J. (2005). *The wisdom of crowds*. Anchor.

Vanlommel, K., Van Gasse, R., Vanhoof, J., & Van Petegem, P. (2017). Teachers' decision-making: Data based or intuition driven? *International Journal of Educational Research*, *83*, 75–83.

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479.

Vitta, J.P. (2021). The functions and features of ELT textbooks and Textbook analysis: A concise review. *RELC Journal*. Epub ahead of print 25 August 2021. DOI: 10.1177/0033 6882211035826

Vitta, J.P., Nicklin, C., & Albright, S.W. (2023). Academic word difficulty and multidimensional lexical sophistication: An English-for-academic purposes-focused conceptual replication of Hashimoto and Egbert (2019). *The Modern Language Journal*, *107*, 373–397.

Webb, S.A., & Chang, A.C.-S. (2012). Second language vocabulary growth. *RELC Journal*, *43*, 113–126.

Zarco-Tejada, M.A. (2019). Automatic profiling of L2-simplified texts: Identifying discriminate features of linguistic proficiency. *Digital Scholarship in the Humanities*, *34*, 661–675.

Zhang, S., & Zhang, X. (2020). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*, 696–725.