Vocabulary Learning and Instruction
Volume 10, Issue 2, December 2021
http://vli-journal.org

Discussion Paper: Using Statistics to Solve Practical Vocabulary Problems

Jenifer Larson-Hall
University of Kitakyushu

Most of us who do research on language acqusition have had to use statistics to evaluate the results of experiments. Some may use only the statistical procedures they learned in graduate school and may thus miss out on new advances in statistics that might shed light on some problems in a more straightforward way. The three papers that conduct empirical studies that I will discuss today have used statistical procedures that you may not be very familiar with—bootstrapping, Monte Carlo simulations, and Rasch (or item response theory [IRT]) analysis. Their use of these procedures, however, means that they are able to give quite precise and interesting answers to the questions that they have asked. The fourth paper I will discuss is not an empirical study but a review of studies and call for future research going forward.

I'd like to start with the paper by Stewart, McLean, and Batty (2021, issue 10.2) entitled "Correlations of modalities of written vocabulary knowledge to listening and reading proficiency: A comparison." In this article, the authors basically used the data that were described in more detail in a separate article published by all three authors in Language Testing (McLean et al., 2020). In that 2020 study, they examined which of the four modalities of vocabulary tests best correlated to reading ability as measured by Test of English for International Communication (TOEIC) Reading section scores. In the present (2021) study, the authors used the data to examine the correlations to listening ability, as measured by the TOEIC Listening section scores and compared that to the reading correlations. Putting aside the question of whether the TOEIC tests accurately measure reading or listening ability,[1] I would like to examine the specific way in which both papers by these authors answered the research questions they set out.

Both of these studies used interesting methodological and statistical methods to significantly increase their power to give interesting answers to their respective research questions. The McLean, Stewart & Batty (2020) data (I will name the data MSB based on the authors' last name order in the 2020 article) consists of 4,000 data points per person, with 103 participants. Each participant gave their answers regarding the third most frequent band of 1,000 words from the New General Service List (Browne et al., 2013), which includes words like *supplier*, *personnel*, and *stimulus*.[2] The test takers answered 1,000 questions about their knowledge for each of the four modalities of tests that have been used to measure vocabulary knowledge (see Table 1).

The first innovative methodology of the MSB data is that it is complete, at least for one 1,000-word vocabulary band. Because the range of possible

Please cite this article as: Larson-Hall, J. (2021). Discussion paper: Using statistics to solve practical vocabulary problems. *Vocabulary Learning and Instruction, 10*(2), 101–113. https://doi.org/10.7820/vli.v10.2.larson-hall

Table 1. Four modalities to test vocabulary as cross-tabulated by meaning vs. form and recognition vs recall

Modality		Example for native English speaker learning Japanese	Example for native Japanese speaker learning English
Meaning recall	See L2 form, write L1 word	I ride a 自転車.	親切なstudentです。
Form recall	See L1 form, write L2 word	We count them every day.	進行は遅い。
Meaning recognition	See L2 form, choose L1 meaning	楽曲 a) game b) movie c) song d) yard	Piece of music: a) 遊び b) 映画 c) 歌 d) 庭
Form recognition	See L1 form, say whether you know it	Check any words you know: □ 遊び □ 映画 □ 歌 □ 庭	知っている単語をチェック： □ game □ movie □ song □ yard

vocabulary words is so large, most studies simply sample from the vocabulary range that they are interested in. This is the first study I have heard of which asks the participants to judge every word. The authors used their complete information to create samples of each 1,000-item dataset in differing lengths to answer the question of how long a sample had to be to accurately reflect the true score of the participants. Because they had not sampled the participants but instead exhaustively tested them, they knew their true score and were thus able to look at what sample lengths were accurate.

The second innovative step was to use bootstrapping with this data. This was not necessary to answer their research question. For a sample of length 100, for example, the authors could have just randomly drawn out 100 numbers from the 1,000 available numbers for each participant to create an appropriate sample. They could have done this for all 103 participants and then correlated the average score of each sample to the own participant's TOEIC reading and listening scores.

However, the authors noted that to find very small effect sizes they would need to sample several hundred students (McLean et al., 2020). Obviously asking even 103 participants to judge 4,000 items must have been a Herculean task, so there is no way to criticize the authors for not having more participants. They thus decided to use another method to increase the generalizability of their data: bootstrapping. Bootstrapping is one recent statistical tool that has been introduced in the SLA field (Larson-Hall & Herrington, 2009) as a way to deal with smaller samples and also a way to overcome the problem of the fact that for the small sample sizes (< 50) used in our field it is essentially impossible to determine whether the data are in fact normally distributed. However, to perform parametric statistics one must assume that the data are *exactly* normally distributed and not slightly heavier in the tails of the empirical (sampled) distribution than the normal distribution, which is called a *contaminated normal distribution* and which research has shown can cause type II errors (Tukey, 1960; Wilcox, 2001). This results in a classification of a difference or relationship as non-statistical when it is in fact statistical.

Bootstrapping treats the data of each participant as a pool of data to be randomly sampled from and then creates an empirical distribution from this random sampling of the data. In other words, bootstrapping does what the researchers themselves would like to do: repeat the experiment. In this case, there was quite a large pool to sample from given that each participant had 1,000 data points. Thus, for the length of 100 items, the authors randomly sampled 100 items from those 1,000 data points, and they did that 1,000 times. Each time a number was randomly picked from the pile of 1,000 data points, it was tallied and then dropped back into the pile of 1,000 numbers--this is called sampling with replacement. Each of those 1,000 samples of length 100 was then reduced to its average number. This was repeated for all 1,000 samples. The authors now had an average score for a sample of length 100 for each participant that was itself the average of 1,000 samples of size 100 from the original 1,000 data points. If this seems confusing, perhaps an illustration will help (see Figure 1). This average, then, was correlated with the TOEIC listening score.

In the McLean et al. (2020) article they had 21 different lengths of tests, and they created 1,000 bootstrap samples from each test (21 × 1,000 = 21,000) for the 4 different types of tests (4 × 21,000=84,000 tests) for each person. There were 103 participants so the total number of samples they had was (84,000 × 103 = 8,652,000). They could then calculate a mean score of the samples for each modality of test at each of the 21 lengths of test. This number represented the scores of the 103 participants but with 1,000 samples from each, meaning it was much more accurate than simply taking a random sample of that length from each participant would have been. Essentially it was like surveying 1,000 × 103 = 103,000 participants for each length in a certain modality.

Now there was nothing special about the MSB data that meant they needed to use bootstrapping. The point I want to bring up here is that anyone can use bootstrapping with their data as a way of generating more resilient and accurate samples.

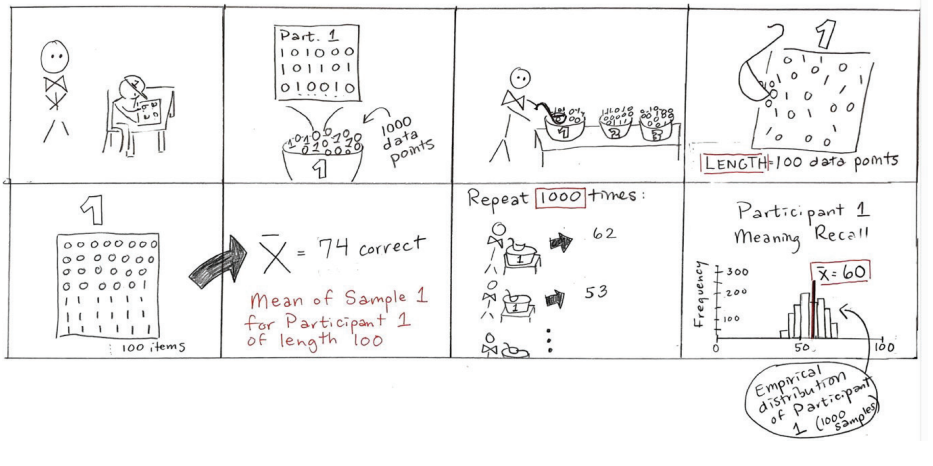

Figure 1. Bootstrapping, illustrated for one participant in one modality with samples of length 100 and 1,000 bootstrapping samples taken.

The current paper used exactly the same methodology as the McLean et al. (2020) article for dealing with the data, the only difference being that the scores were correlated with the TOEIC listening test.

This paper did not show the data graphically so I just wanted to show what that correlation between the vocabulary levels test at length 100 and the TOEIC listening data looked like (Figures 2 and 3).[3] By the way, to plot Figure 2 I used Mizumoto's langtest.jp site (http://langtest.jp/shiny/cor/) and the ellipse shows the robust part of the correlation. I included Figure 3 because it has a LOESS line which follows the pattern of the data instead of just drawing the straight least-squares line on the data.

The strong correlation in this data is clear. The point is that this data looks like every other scatterplot. There's nothing strange about the MSB data even though it was bootstrapped.

How does one perform this bootstrapping? McLean et al. (2020) said they used an Excel formula. The formula would have sampled with replacement X number of items from the pool of 1,000 and then calculated an average for that one sample, then repeated that 999 more times. Simple bootstrapping is not difficult. Bootstrapping can be used together with other robust procedures, however. One discussed in Larson-Hall and Herrington (2009) is trimming. This removes a certain percentage of the data on both ends of a distribution, thus excluding outliers in a principled manner and being more robust to single extreme values than the mean score but keeping more of the data than the median, which throws away all but one or two values.

The fact of the matter is that I cannot stop praising this paper for its use of larger samples to begin with and then for using bootstrapping to ensure that the

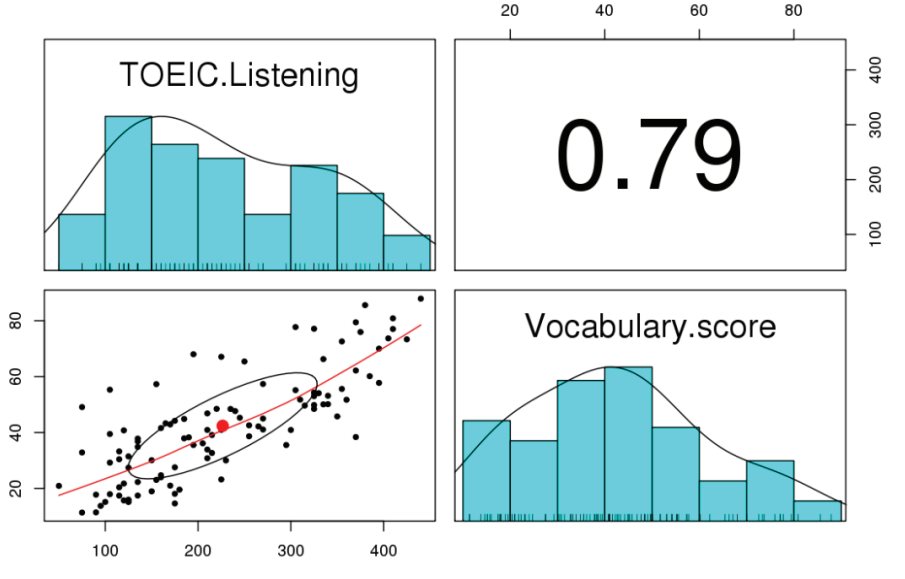

Figure 2. McLean et al. (2020) data for 100 samples in the meaning-recall modality.

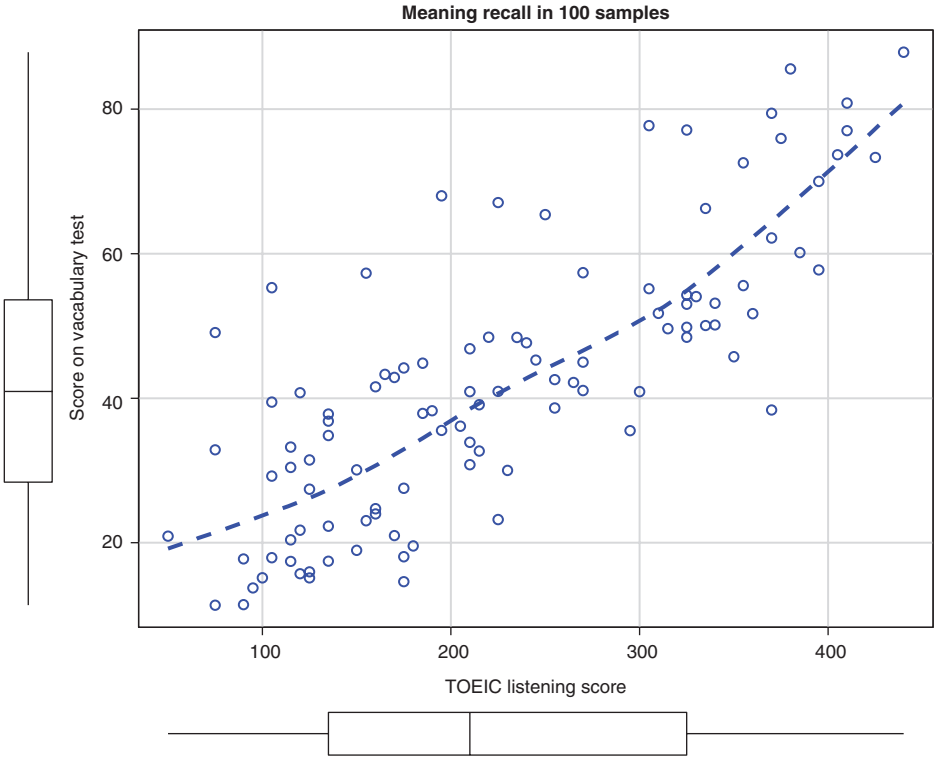

Figure 3. Correlation between scores on the vocabulary test (sampled at length 100) and the TOEIC listening test for N = 103 participants.

samples would be highly representative of the population from which they are drawn. Tversky and Kahneman (1971) point out that most people have wrong intuitions about samples and think that all samples are similar to their population. For example, if a researcher found a statistical correlation of $r = 0.30$ in a sample of $N = 40$ participants, most researchers think that testing $N = 20$ participants would similary result in finding a statistical correlation of about the same magnitude. Tversky and Kahneman call this a "belief in the law of small numbers" when what we can actually believe in is the representativeness of large numbers and large samples only.

The McLean et al. (2020) paper also showed pretty convincingly that meaning recall tests (see L2 word form and write L1 word) are the best types of tests to use when one wants to test vocabulary ability. Meaning recall tests had the highest correlation to reading proficiency, reaching $r = 0.76$ at a test length of 40 items, indicating that vocabulary ability as measured by meaning-recall explains almost 50% of the variance in the TOEIC reading portion. The superiority of the meaning recall tests to the other three types of tests still held true in a comparison involving the same amount of time spent on the test (by the participant). And it turned out this result also applies to correlations with vocabulary skills and the TOEIC listening tests as well.

Of course, meaning-recall involves more work for the person scoring the test, but McLean has been working on reducing that burden by creating a website where the computer can learn what answers you favor and ask you to judge unorthodox answers. This is of course the vocableveltest.org site described in McLean, Huston, Raine, Kim, Ueno, Pinchbeck and Nishiyama called "The internal consistency and accuracy of automatically scored written receptive meaning-recall data: A preliminary study."

To me it seems that Stuart McLean has organized a fruitful and practical line of research that he is methodically pursuing in collaboration with many others. With strong research showing that meaning-recall tests (see L2 word form and write L1 word) are the best way to measure vocabulary that is useful for reading and listening abilities, he has set about trying to make such tests more user friendly. I had known about McLean's vocableveltest.org site before, but what I hadn't realized before was how it related to best practices for vocabulary testing as uncovered by research.

I'm struck by the research result, quoted in the paper as coming from McLean et al. (2014), that even Japanese students at more exclusive schools (with a *hensachi* over 61)[4] did not demonstrate mastery of the first 1,000 words of English. It seems clear that at least in Japan, the vocableveltest.org might be fruitfully used by almost all university professors who are teaching English as a way to discover where their students might have gaps in their basic knowledge of English. Since bands as small as 100 can be used, teachers have an easy way to test their students' knowledge of foundational English vocabulary a little at a time.

I was also impressed with the measures implemented at this site to prevent cheating. I think that is an important consideration these days when online dictionaries can be easily accessed during testing and when some students may be doing testing at home, not under our watchful eyes.

Now the McLean et al. paper does something very similar to the bootstrapping in the Stewart, McLean, and Batty paper: it uses Monte Carlo simulations to determine how many items will be sufficient to accurately represent a learner's knowledge of the vocabulary in a 1,000-word band. Basically, Monte Carlo simulations use the same process as bootstrapping, although instead of gathering data from live participants, a computer randomly simulates a distribution whose mean is specified in advance. Thus, for Figure 9 of McLean et al. (2021, issue 10.2), the computer creates a normal distribution of numbers whose mean is $X = 750$ with 1,000 samples of the distribution. The researcher then randomly samples some number of 5, 10, 20, and 100 item samples from that distribution. From Figure 9, it looks like 1,000 samples were taken at each of those test size lengths. The average score of that sample is then calculated, and that is plotted for the number of times each of those possibilities is chosen. Thus we can see that *when we actually know* the true score if we only sample five items from the entire distribution (and this would be equivalent to testing five items from a 1,000-band vocabulary level), we could end up with an average score of 200 points occasionally (maybe 10–15 times). Most often we would get a score of around 800 (400 times) but basically with so few points, we could end up almost anywhere along the continuum! Five points are therefore not very likely to get us close to the real score. Obviously 100

sampled items get us within a much closer range to the real score although it may slightly under- or over-estimate the real score. Figure 10 shows that sampling 50 items is pretty close to 100 items although not quite as accurate. Of course, since the McLean et al. (2020) paper went through the same process with actually sampled data, meaning we knew each participant's real score, it was already shown that a sample size of 40 was almost as reliable as a sample size of 100 and the size that the authors seemed to consider "good enough."

Does this mean that the McLean et al. (2021) study did not actually require the Monte Carlo simulation to prove its point? I'm pretty sure it does! I think it could have used the MSB data set. However, does this mean that the McClean, Stewart, and Batty (2020) as well as the Stewart, McLean, and Batty paper in this volume did not actually need to sample students and could have just as well used a Monte Carlo simulation? Well no, because the authors did want to find out exactly how the four test modalities would compare to each other, and they couldn't tell the computer what the mean scores for each modality were in advance the way the Monte Carlo simulation could.

I do quibble a bit with the McLean et al. (2021) paper when it asserts that "Figure 11 suggests that even samples of 100 or 200 items can occasionally result in inaccurate estimates." I assume this refers to the fact that although the true mean is 750, the sampled average with 100 items could be as low as 630 or as high as 870. Statistics are probabilistic and whenever we calculate some statistic from a sample, be it the vocabulary knowledge of a 1,000-band set of words or the effect size of a *t*-test, our statistics may return a point value but any point value is only an estimation and will never be the true score for the population. If that point is understood, then confidence intervals (CIs) become much more valuable. CIs will give an interval around the point value that could plausibly contain the true value, with 95% confidence. Of course, sample sizes of 100 or 200 will have smaller CIs than sample sizes of 20 or 50 but that doesn't mean they are inaccurate. *No* statistic is *accurate*, meaning exactly right. As Crawley (2012) says, all models are wrong, but some models are better than others (p. 403). Statistics can give us the best wrong model, but it will always still be wrong.

I have to say that I have heard Stuart discuss his vocableveltest.org site several times and because of this paper I actually went and tried it out myself. I am doing vocabulary activities with several classes I teach, but I'm afraid the website isn't useful to me as is. I am using Gardner and Davies' (2014) Academic Vocabulary List at my university, and that list is not yet available in vocableveltest.org. Another class I teach has vocabulary activities with the words in the textbook, and I don't see any way to enter your own list of words into vocableveltest.org yet. Thus, unless I want to test students' abilities with basic vocabulary using the NGSL I wouldn't use this tool yet. I also tested out the site and found that some of the sentences are problematic. For example, I used the site to test out my Japanese ability by answering form recall items (see L1 form, write L2 word; see Figure 4). As you can see, my Japanese ability in this test was not great, but notice the second sentence, "Are they pair?" This is not correct English and if I got this strange sentence with a sample of only five sentences it makes me wonder what percentage of sentences might be problematic.

Prompt	Your Answer	Point	
We **count** them each day.	数える	✓	
Are they **pair**?	二つの	✗	Show Answers
They often **invite** a few friends to come over to their house.	招待する	✓	
They **deliver** the food.	届く	✗	Show Answers
progress is slow.	進化	✗	Show Answers

Figure 4. Results from the vocableveltest.org for a form-recall test (done by a native English speaker).

However, I have to say I am extremely glad to see that this website has been created. I really enjoyed the format where I could see my incorrect answer and then see possible correct answers too. I do think the site could be useful for large numbers of researchers. I imagine, since the last line of the paper states it, that the website will continue to be improved and that McLean is open to adding new types of data to the website so researchers like me could use our vocabulary lists too.

The review paper by Kim, entitled "Considerations and challenges in longitudinal studies of lexical features in L2 writing (2021, issue 10.2)," indicates that this author also has an amibitous agenda that could help researchers. It would be great if there were eventually a website where researchers could upload a piece of L2 writing or transcribed speaking and get information about the lexical sophistication, diversity, density and accuracy of the vocabulary used in that writing. Actually, as Kim notes, for English that day may not be far away, as the Crossley, Kyle, and colleagues' tools (TAALED, TAALES, GAMET) can implement algorithms to check for many of these things.

I have examined these tools in my own research, however, and I have found that they are not as helpful when the target language is not English. I had a case study with five L1 English speakers learning Japanese. I tested the participants over 3 years and transcribed their storytelling of Japanese picture books. Previous research of this type has simply reported on the type/token ratio of such utterances, but I thought that with the more sophisticated tools available today I would be able to examine lexical density or sophistication and track how that changed over time. However, the tools are not created for dealing with Japanese and although I was corresponding with Scott Jarvis for a few months as I tried to jury-rig some of the tools to work for Japanese, including learning how to manipulate Jarvis' Python codes, it was slow work and I eventually gave up. This, of course, is an additional challenge if we assume that the L2 writing Kim alludes to includes any L2 other than English.

Beyond this problem, as Kim notes, there is still controvery surrounding the best ways to measure and compute various ideas like lexical sophistication, lexical errors, or any of the other measures of L2 writing. For example, on the

TAALED website page (https://www.linguisticanalysistools.org/taaled.html), there is a paper cited in 2021 that looks at the minimum text length necessary for reliable lexical diversity measures. In other words, it is not even clear yet how long a text should be to reliably measure its lexical diversity.

Kim notes that if vocabulary development is investigated through writing (or I would add, transcribed speaking measures), then it would be better to look at writing over spans longer than 1 year where participants are in a situation where they are likely to get large amounts of input. Since language acquisition is a rather slow process if we are really thinking about real productive or receptive abilities and not just scores on a test where time is available for explicit analytical abilities to come into play, then this recommendation is probably a great one for any researcher who wants to study development at all! Frankly, I've come to think that studies that test out a teaching technique over a short period (like an hour!) and look at the results are usually fairly worthless. We don't have any evidence that groups that might perform better on an immediate posttest, or even at a delayed posttest of several weeks later, will still retain any of that information better over the long run. I know that longitundinal research of the type Kim is advocating and which I am seconding takes a lot of time and there's always a lot of attrition from participants, but better to struggle with that than waste our resources with short-term studies that aren't worth the paper they're printed on!

The issue I see with Kim's proposal, however, is that this area of research looks enormous and I am not sure it is at a state where many concrete recommendations can be made. Obviously I agree that examining vocabulary development over the long run when participants are getting lots of input is a great idea. However, Kim herself states, "we do not know much of L2 lexical development, and more research is needed" (p. 3). The papers cited in the literature review certain do not paint any kind of vivid picture of what is going on with lexical development. When researchers in the field are still investigating what lexical features to measure and how to best measure them, and when in fact very little research has actually been done longitudinally, this call for more research sounds too broad to my ears.

Let's take just one area that Kim mentions. She says that a consideration in the research design is what the basic unit of lexical analysis is and lists possible units as being tokens, lemmas, or word families. However, this area alone is quite controversial and could probably encompass a substantial research agenda. The unit of flemma, explained by the McLean et al. paper given in this session as "a base word form and inflectional forms, regardless of POS" (p. 6), differs from the lemma by not separating words with different parts of speech that are identical in form into separate categories (e.g., use_N and use_V are separate lemmas but one flemma). However, the question of whether the word family, lemma, or flemma is more appropriate for use with English learners is something that has, to my mind, still not been resolved and has been addressed by several researchers from our vocabulary SIG (McLean, 2017; Stoeckel et al., 2020). I want to turn back to the safety of the other papers here, which I think address very small but practical and solvable issues.

Therefore, I would like to mention Nicklin's paper, called "Developing a measure of proper name familiarity for Japanese university students (2021, issue 10.2)." This very small and discrete study seems to be one brick in an agenda determined to understand whether proper nouns disrupt comprehension abilities. I have to admit that before reading this paper I would have definitely said that they did not, but it appears there is at least as much support for the conclusion that proper names make reading English difficult as there is for the finding that studying vocabulary in semantic sets is worse than studying it in thematic sets. That very specific vocabulary finding based on only a few studies got a whole chapter in Folse's Vocabulary Myths book (Folse & Briggs, 2004), while I have never heard of proper noun problems before, so I am glad Nicklin has drawn my attention to it.

Nicklin does not yet provide a paper that establishes whether proper nouns impede comprehension but rather provides an example of a tool that he will use in his investigation of this question. He thus provides an example of how to proceed in other endeavors where new tools are created. These examples are needed! I remember in my early research I used a grammaticality judgment test that had been used in critical period studies. In my own research, I cited the reliability statistics of that test from the previous research papers as if the high reliability numbers found in someone else's paper meant that the test was reliable without any further demonstration on my part. What I didn't understand was that reliability is a function of how a particular sample performs on a particular test, so that the only reliability statistics I should have provided were those for my own sample.

Nicklin here uses Rasch modeling, or what is known as item response theory (IRT), to validate what ultimately becomes a range of 30 items that span a continuum of familiarity for Japanese university-level learners of English. Again, just as the first two papers I mentioned used some innovative statistical methods, this paper also uses a statistical method that may not be very familiar to some readers. Although IRT is by no means a new procedure it is not available in the base version of SPSS and possibly because of this is not widely understood.

Paolillo (2000) notes that those who ignore methodological concerns do so at the peril of misinterpreting their data and making false conclusions about their experiments. I think Nicklin has clearly shown that this could have easily happened in his study if he had not investigated the ability of his Likert-scale familiarity questionnaire to distinguish four levels of familiarity. Other researchers often use Likert-scale questions without any analysis of whether the scale is actually reliable or valid, and again, that is why this paper is quite exemplary.

IRT improves upon classical test theory. The R packages that Nicklin used to conduct his analysis were called TAM (test analysis modules) and eRm (extended Rasch measurement). Using such packages one can call for the descriptive statistics that are calculated in a classical test analysis approach to the data: item facility (which calculates how many people got the item "correct" out of the people who took it), item discrimination (a measure that lets the researcher see whether those who scored highly overall on the test scored highly overall on a particular item), and the biserial correlation with the item excluded. While classical

test analysis basically examines how difficult items are for the test takers, IRT gives information about the skill levels of the test takers themselves. While classical test analysis cannot talk about reliability beyond the scores of the test takers themselves and cannot generalize test scores beyond the sample and items tested, IRT's advantage is that it is parameter invariant, meaning that "item statistics that are obtained from the application of IRT models are independent of the sample of examinees to which a test is administered" (McKinley, 1989). It is also an advantage that it can score individuals according to their ability levels and give error measurements for individual items in the test. I've always been intrigued by the promise of IRT for doing adaptive testing. Because it can take the responses of a test taker and see what scores they receive on items at different difficulty levels it is able to quickly discriminate what a test-taker's ability level is.

Nicklin's analysis was more interested in the items of the test and which items would span a range from very unfamiliar through very familiar proper nouns. I checked out Nicklin's supplementary materials and while I cannot claim that I could follow the entire analysis, one great thing about R is that all of the R commands that Nicklin used, including his data and commands for drawing the graphics found in the paper, are reported online and available to anyone who wants to check them or use them to do something similar to what Nicklin did. If you are not familiar with the free R statistical program yet, I cannot recommend highly enough that you become so. It is the statistics of the future.

In conclusion, three of the studies presented today resemble each other in that they have used lesser-known but powerful types of statistical analysis to make their conclusions more valid and reliable. Although it may seem that such analysis is too sophisticated or advanced for some readers, with an initial investment of time to learn the basics of the R statistical language they are not beyond the abilities of normal researchers. I do want to acknowledge that Kim is no stranger to statistical analysis herself, as shown by her careful work in Kim et al. (2017), where she used Principal Components Analysis to find 12 components of a lexical sophistication measure and then used correlations and regression between the components and a measure of writing proficiency to discover which components could successfully predict writing proficiency. From these papers, we can see the benefit of learning about the latest statistical methods and employing them carefully in our own research.

Notes:

1. Nicholson (2015) says that although the TOEIC scores have been found to be reliable some researchers question the premise of using the TOEIC test in such situations at all since the test was developed to measure how well English L2 users can communicate in a business workplace. Nicholson says that although the TOEIC test is fairly prestigious in Asia and is used by many employers, the lack of independent verification that the TOEIC measures actual reading or listening skills is "startlingly low" (p. 225).

2. I used the document entitled "NGSL + 1.01 + with + SFI" and the words listed between 2001 and 3000 to obtain these words. McLean, Stewart, and Batty (2020) said that they used flemmas as sorted by the words' SFIs (standard frequency indices), but the Excel file column itself specifies that these are lemmas so I do not know whether anything was done to the original document to change to a flemma distinction in words.
3. Note that the authors provided me with the data upon request. My belief is that this should be the norm for any published paper. One step further is to make your data publicly available on a site such as github (as Nicklin did), Open Science Framework repository (osf.io), or the IRIS database (https://www.iris-database.org/iris/app/home/index;jsessionid=9B0A42048B0D992898D00BC52905D123).
4. The Japanese term hensachi when used in connection with universities refers to the average score of students entering any university based on a national test whose scores are normed to a scale where 50 represents the mean score of all students who took the test. Students whose scores were one standard deviation above the mean would receive a 60. By the 68–95–99 rule for a normal distribution then, 68% of all students would fall within one standard deviation of the mean, so only 16% of students (32%/2) would have scored above 60.

References

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for. *Vocabulary Learning and Instruction, 3*(2), 1–10. https://doi.org/10.7820/vli.v03.2.browne

Browne, C., Culligan, B., & Phillips, J. (2013). *The new general service list*. https://www.newgeneralservicelist.org/word-list-research

Crawley, M. J. (2012). *The R book*. John Wiley & Sons.

Folse, K. S., & Briggs, S. J. (2004). *Vocabulary myths: Applying second language research to classroom teaching*. University of Michigan Press.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*(3), 305–327. https://doi.org/10.1093/applin/amt015

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal, 102*(1), 120–141. https://doi.org/10.1111/modl.12447

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368–390. https://doi.org/10.1093/applin/amp038

McKinley, R. L. (1989). Methods, plainly speaking: An introduction to item response theory. *Measurement and Evaluation in Counseling and Development, 22*, 37–56. https://doi.org/10.1080/07481756.1990.12022910

McLean, S. (2017). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*, 823–845. https://doi.org/10.1093/applin/amw050

McLean, S., Hogg, N., & Kramer, B. (2014). Estimations of Japanese university learners' English vocabulary sizes using the vocabulary size test. *Vocabulary Learning and Instruction, 3*(2), 47–55.

McLean, S., Stewart, J., & Batty, A. O. (2020). Predicting L2 reading proficiency with modalities of vocabulary knowledge: A bootstrapping approach. *Language Testing, 37*(3), 389–411. https://doi.org/10.1177/0265532219898380

Nicholson, S. J. (2015). Evaluating the TOEIC® in South Korea: Practicality, reliability and va-lidity. *International Journal of Education, 7*(1), 221–233. https://doi.org/10.5296/ije.v7i1.7148

Paolillo, J. C. (2000). Asymmetries in universal grammar: The role of method and statistics. *Studies in Second Language Acquisition, 22*, 209–228. https://doi.org/10.1017/S0272263100002035

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics, 41*(4), 601–606. doi: 10.1093/applin/amy059

Tukey, J. W. (1960). A survey of sampling from contaminated distributions. In I. Olkin (Ed.), *Contributions to probability and statistics: Essays in honor of Harold Hotelling* (pp. 448–485). Stanford University Press.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105. https://doi.org/10.1037/h0031322

Wilcox, R. R. (2001). *Fundamentals of modern statistical methods: Substantially improving power and accuracy* (Vol. 249). Springer.