Using dictionaries to estimate vocabulary size: essential, but rarely followed, procedures

Paul Nation *Victoria University of Wellington*

This article describes the steps that need to be followed when sampling from a dictionary to make a test of vocabulary size. Because these steps have not been followed, most published estimates of vocabulary size are misleading.

One of the earliest published investigations of vocabulary size was Kirkpatrick's (1891) study of his own knowledge of the words in Webster's unabridged dictionary. Since that time there have been numerous attempts to estimate the vocabulary size of native speakers using dictionaries as starting points. There is little point in providing a chronological account of the research because, with one or two notable exceptions, researchers were unaware of the relevant previous research. This at best resulted in the rediscovery of the same principles and, at worst and most often, resulted in misleading estimates of vocabulary size.

The aim of this article is to describe the procedures that should be followed when using a dictionary to provide the basis for an estimate of vocabulary size. This will be done by examining the few studies that suggested the procedures and then by listing and describing them. The basic question answered here is 'How can we take a representative sample from a dictionary?'

Group-administered measures of vocabulary size have involved taking a random sample of words from a dictionary, usually by taking the first or *n*th word on every *m*th page. This random sample is a proportion of the dictionary. In the Goulden, Nation, and Read (1990) study, the sample was one one-hundredth (1%) of the 129 000 word dictionary. A person's score on a test based on the sample is multiplied by the proportion (100 times in the Goulden, Nation, and Read study) to gain an estimate of total vocabulary size.

I Thorndike's review

The earliest and most important study in this area is 'The vocabular-

Language Testing 1993 10, 1
27 – 40

ies of school pupils' by Edward Thorndike (1924). Thorndike reviewed nine studies of vocabulary size, finding a wide disparity in their estimates. He argued that this disparity was largely the result of the effect of the sampling method used, combined with dictionary size. In his review he described four procedures that should have guided subsequent research. First, he showed the importance of using a sample source, a dictionary or frequency list, that did not exclude words likely to be known by the people being investigated. Any vocabulary study that used a sample source that was too small would underestimate vocabulary size.

Secondly, Thorndike showed that it was necessary to have clearly stated criteria regarding what was to be included in a word family. For Thorndike, those criteria were the same as those to be used for the Thorndike and Lorge (1944) word-frequency list. Thus, most words with derivational suffixes were considered to be different words from the base form and other derived forms. As we shall see, failing to include derived forms in a word family will bias the sampling towards better-known items. Most critical, however, was the decision regarding subsequent homographs. Many words, particularly high-frequency words, have several entries in the dictionary. *Draw* has two entries in Webster's *Third*. Because there are several entries there is more likelihood of such forms being chosen in the sampling.

Any dictionary-based vocabulary size study that did not have well-thought-out criteria would be leaving the decisions to the dictionary makers. This would usually result in an over-estimation of vocabulary size.

Thirdly, and most importantly, Thorndike demonstrated that it was necessary to use a sampling procedure that did not bias the selection towards high-frequency words. Procedures that involved choosing the first word on a page regardless of whether it was the first full entry and whether it was a subsequent homograph were particularly susceptible to this bias. This bias occurred simply because high-frequency words occupied more space per entry and had more entries than low-frequency words. The greater the size of the dictionary, the more the space given to high-frequency words, and thus the greater the overestimation of vocabulary size. Thorndike suggested two ways of overcoming this bias. One way was to number all the entries and select every *n*th word. 'To take a random sample of 100 words from a dictionary containing, say, 28,000, we should take words 1, 281, 561, 841 etc., or words 2, 282, 562, 842 etc., or the equivalent' (1924: 73). Another way is to select the first word whose definition *begins* on that page, as long as subsequent homographs are allowed for (Thorndike, 1924: 73, footnote 1). Unfortunately, this way could introduce bias because of the higher selection probability for words on pages containing few words. Such a procedure is a one-stage cluster sampling with one item (word) sampled per cluster (page). In order to avoid bias, clusters need to be the same size. Studies that use procedures that introduce bias will greatly overestimate vocabulary size because their samples will contain too many high-frequency words which will be known by most people sitting the tests.

Fourthly, Thorndike demonstrated a way of checking the representativeness of the sample. He used the 1921 edition of *The teacher's word book*, a word-frequency list, to see if the samples used in Gerlach's (1917) and Brandenburg's (1918) studies contained an appropriate number of words from the various frequency levels. He found that both samples were heavily biased towards high-frequency words.

Thorndike found that none of the nine studies published between 1907 and 1919 that he reviewed was methodologically sound. They failed on three and sometimes on all four of the procedures he suggested.

These four procedures were sufficient to provide representative samples for vocabulary-size research, but unfortunately Thorndike's review was published in a collection of papers rather than in a more easily accessible journal. Indeed, the first reference to this 1924 review is in a 1953 study by Bryan, which refers to the review but fails to make use of its findings. The second reference to the review is in a 1963 review by Lorge and Chall. Lorge was a colleague of Thorndike and his review made use of Thorndike's suggestions.

II Later reviews

After Thorndike, the next article to address the methodology of the measurement of vocabulary size sensibly was Williams (1932). Williams compared a sample drawn from a dictionary with an independent list of children's vocabulary. He found that the number of words from the children's list in the dictionary sample was much greater (by 45 times) than it should be. Apparently, Williams was not aware of Thorndike's review. However, he reached the same conclusion that the spaced sampling method combined with dictionary size resulted in an overabundance of high-frequency words in the sample. The two sampling procedures he described to overcome this bias were the same as those described by Thorndike eight years earlier. Williams also pointed out that using abridged dictionaries could result in an underestimation of technical or specialist vocabu-

laries. Williams thus rediscovered three of the four procedures described by Thorndike.

Lorge and Chall (1963), in a critical analysis of the widely described Seashore and Eckerson (1940) study, elaborated on the remaining procedure of setting up clearly described criteria for deciding what was to be counted as a word. They argued that proper names, geographical place names, word parts, and names of flora and fauna should be excluded from measures of vocabulary size. They found that the Seashore and Eckerson test greatly overestimated vocabulary size because it did not apply three of the four procedures suggested by Thorndike. Seashore and Eckerson did not set up sufficient criteria for what was to be counted, but relied on the dictionary makers' decisions. They did not use a suitable sampling procedure and did not use an external check on the representativeness of their sample. Words with multiple entries were a particular influential source of bias towards high-frequency items in their sample. Seashore and Eckerson were aware of Williams's (1932) study but failed to see its significance for their study.

Seashore and Eckerson, however, did make use of an internal reliability check on their sampling, which was a useful addition to the four procedures described by Thorndike. Seashore and Eckerson compared proportions of parts of speech in subsamples to see if they were similar.

Unfortunately, the most perceptive critics of attempts at making dictionary-based tests did not make their own and it has been left to others to put their suggested procedures into practice. It is sufficient to say that none of the 20 or more published studies of vocabulary size since 1907 has made full use of the necessary procedures. They have thus produced misleading estimates.

Over the last 10 years there has been a considerable revival of interest in vocabulary size and knowledge, partly as a result of argument over the feasibility and effectiveness of direct vocabulary teaching. This interest has led to research which allows us to build on the procedures described by Thorndike, and the rest of this article draws on this research to restate, add to, and refine the sampling procedures that a proper dictionary-based study of vocabulary size must follow.

III Procedures for drawing a representative sample from a dictionary

1 Choose a dictionary that is big enough to cover the known vocabulary of the people being investigated

This does not necessarily mean that an unabridged dictionary must be used when studying the vocabulary size of educated adult native speakers. Recalculations of methodologically faulty studies (Thorndike, 1924; Lorge and Chall, 1963) and recent more methodologically sound studies (Goulden, Nation, and Read, 1990; D'Anna and Zechmeister, forthcoming; Nation and Ellis, in preparation) indicate that educated adult native speakers' base word vocabulary is around 20 000 words. So, for educated adults, a dictionary should contain at least 30 000 base words. This is easily within the range of abridged desktop dictionaries such as *Collins English dictionary*. D'Anna and Zechmeister (forthcoming) checked the inclusiveness of the *Oxford American dictionary* by comparing it with two advanced word lists. Goulden, Nation, and Read (1990) supplemented Webster's *Third* by using recent addenda.

2 Use a reliable way of discovering the total number of entries in the dictionary

In order to make calculations of vocabulary size it is necessary to know the total number of entries in the dictionary, or the total number of the particular types of words, such as base words, that are the focus of the study. Dictionary makers are not greatly concerned about the number of entries in a dictionary except to the extent that this can be used to advertise the dictionary and increase its sales. Not surprisingly, publishers' statements exaggerate the number of entries. For example, Webster's *Third* claims to have 450000 entries. Two independent studies (Dupuy, 1974; Goulden, Nation, and Read, 1990) reach the same figure of 267000. *The shorter Oxford English dictionary*, in a computer-based count, was found to have 73582 entries (Dolby and Resnikoff, 1967). The note on the dust-jacket, however, says 163000.

Diller (1978) accepted the Webster's *Third* figure of 450000 without checking and, because the number of words in the testing sample in proportion to the total number of entries was the basis for his calculations, reached wildly inflated estimates of vocabulary size.

The total number of entries can be found by counting each entry manually (D'Anna and Zechmeister, forthcoming) or with a computer (Nation and Ellis, in preparation), or by counting a sample of the dictionary (Dupuy, 1974; Goulden, Nation, and Read, 1990).

3 Use explicit criteria for deciding and stating (a) what items will not be included in the count and (b) what will be regarded as members of a word family

The main reason for excluding items from a sample or an estimation of dictionary size is to avoid the effect that the dictionary makers' inclusion policies can have on the estimation of vocabulary size.

Lorge and Chall (1963:149) consider it 'debatable' and 'highly questionable whether [encyclopaedia and word-part entries] can be considered "words" for estimating vocabulary knowledge' (Lorge and Chall, 1963:151). Encyclopaedic entries include names of historical personages, fictional characters, geographical place names, and the scientific names of flora and fauna. As well as it being questionable whether to include these as words, the major difficulty lies in the arbitrary nature of their inclusion in the dictionary. If geographical place names are included for example, where does the inclusion stop? Will the name of *every* country in the world be included or only those thought to be of interest in areas where the dictionary is sold? If plant and animal names are included, what about the multitude of insects and bacteria? Dictionaries clearly differ on their policies of what will be included and this is not usually stated in a rigorous way.

By excluding types of words where different dictionary makers have different policies and by limiting the estimation of vocabulary size to the types of words that all would agree on as a part of someone's vocabulary, it is possible to make an estimation of vocabulary size that can be generalized beyond the particular dictionary studied. This estimation can then be used for more than relative purposes and can be used as a tool for research into the relationship between vocabulary size, skill in language use, and the acquisition of knowledge.

Another important reason for excluding items from the count is practicality. When estimating vocabulary size, the smaller the ratio of the sample to the whole dictionary, the more sensitive the measurement will be. For example, a sample which is one twentieth of the dictionary is better than a sample which is one hundredth of the same dictionary. However, the smaller the ratio, the greater the number of words in the sample, and the greater the difficulty of testing them all in a reasonable time. It is thus more practical to reduce the size of the dictionary by excluding items. For example, the exclusion of proper names, compound words, and various items like abbreviations, word parts, symbols, etc., from a study of Webster's *Third* reduced the total size by almost 50% (Goulden, Nation, and Read, 1990).

Another reason for having criteria for inclusion is to avoid a bias in sampling towards high-frequency words. There is a reasonably strong connection between frequency of occurrence in the language and likelihood of people knowing the word (Anderson and Freebody, 1983). It is important that words with related forms and meanings are dealt with as word families. This means that either all items except the base form are excluded from the sampling or that the related forms are sampled separately from the base form and are tested in a different part of the test. If this is not done then the sampling will be biased towards high-frequency items because high-frequency items have more related forms than low-frequency items. Clearly, regarding items such as *legal*, *legalese*, *legalism*, *legality*, *legalize*, which are all separate entries in the *Collins English dictionary*, as different words would increase size estimates. Different dictionary makers have different policies regarding related forms. Some list them under the base word, but list homographs with a related meaning but a different part of speech separately. Others list related forms as separate entries. Dictionary makers also differ in whether they include compound words which are made up of items that are already included in the dictionary with similar meanings, such as *coaxial cable*.

Thorndike (1924) regarded a word family as consisting of the base form, and the base plus *-s*, *-ed*, *-ing*, *-ly*, *-er*, and *-est*. Recent research (Tyler and Nagy, 1989; White, Power, and White, 1989) shows that particularly with children aged 10 and above it is reasonable to assume that they are able to derive the meanings of words with common prefixes and suffixes and a known related base form. There are frequency studies of prefixes (Stauffer, 1942; Bock, 1948) which can help in setting up criteria for dealing with such items. Here are examples of derived forms from *Collins English dictionary* that have a clear meaning relationship to their base form and are all separate entries. *Unreal*, *premarital*, *transatlantic*, *misbelief*, *cranial*, *craniate* (having a skull or cranium), *craniology* (branch of science concerned with the shape of the skull), *craniometer*, *craniometry*, and *craniotomy*.

Goulden, Nation and Read (1990) used the following criteria for classifying items as derived words.

> A corresponding base word must occur as a main entry in the dictionary. The meaning of the derived word must be clear from the meaning of the parts that make up the word or involve the minimum of extra learning ... Words consisting of common prefixes ... attached to base words are marked 'derived' (Goulden, Nation and Read, 1990:345).

In a later study (Nation and Ellis, in preparation) all the common

prefixes were listed in the criteria. In both studies, Nagy and Anderson's (1984) scale was used to determine degree of relatedness and provide a cut-off point.

It is also necessary to have criteria for deciding consistently which form will be considered as the base form. Clearly it should be the least inflected form but when two or more forms such as *jurisprudence* and *jurisprudent* are the least inflected forms, then either part of speech, order of occurrence in the dictionary, form of the definition (does the definition of one contain the other form?), or frequency or historical criteria must be used to make a decision. The criteria for doing this should be clearly described. Nation and Ellis (in preparation) used the following criteria:

> Words with a less inflected base form of related meaning in the dictionary were classified as derived words. So *pseudonymous* in the sample was classified as derived because the base form *pseudonym* occurred in the dictionary. Nagy and Anderson's (1984) scale was used to help decide this (Goulden, Nation, and Read, 1990). Where it was difficult to make a decision on the relative degree of inflection, the first entry was chosen. For example, *prophesy* was chosen over *prophet* because *prophesy* occurred first in the dictionary. This rule was not followed [a] if the definition of the first word used the second word as a base, e.g. *puerperal* 'relating to or occurring during the puerperium', [b] if the first word was marked as rare. In these cases the second item was classified as the base word.

Thorndike (1924), Williams (1932) and Lorge and Chall (1963) showed the importance of treating subsequent homographs of a word separately. The failure to have criteria for dealing with these items was a major cause of overestimation of vocabulary size in most studies. As we shall see in the next section, subsequent homographs of related meaning to the first entry must be excluded from the sampling.

A properly conducted sampling can take account of the various types of words in the dictionary. There is no need to have to keep to the policies of exclusion used by other researchers as long as the sampling criteria are clearly stated with examples, and figures are given for the various types. This allows other researchers to make calculations and adjustments according to their own policy.

4 Use a sampling procedure that is not biased towards items which occupy more space and have more entries

When sampling, the criteria for exclusion and inclusion of items can be applied before, while, or after the sampling is done depending on practicality.

Thorndike (1924) suggested using numbered entries. This means counting off every *n*th word and including it in the sample. Subsequent homographs and derived forms would need to be distinguished. Suarez and Meara are preparing a test of Spanish based on a selection from numbered entries.

Thorndike (1924) also indicated that choosing the *n*th complete entry which was not a subsequent homograph on every *m*th page was an alternative procedure. This procedure was used by Goulden, Nation, and Read (1990), but it has problems if simple random sampling is assumed because each page contains a different number of words, and pages with a few words are overweighted in estimates of vocabulary size.

Another approach is to use a random sampling technique. *The English word speculum* (Dolby and Resnikoff, 1967) contains a randomized list of all the main entries in *The shorter Oxford English dictionary* with homographs removed. Nation and Ellis (in preparation) made a computerized random selection.

D'Anna and Zechmeister (forthcoming) used a stratified sampling technique based on the letters of the alphabet. This type of sampling helps overcome bias by making the structure of the sample represent as closely as possible the structure of the population by having the same proportion of words beginning with each letter in the sample as in the population.

The goal of all these procedures is to end up with a representative, nonbiased sample of a manageable size.

5 Choose a sample that is large enough to allow an estimate of vocabulary size that can be given with a reasonable degree of confidence

Surprisingly, the absolute size of a sample has a greater effect on the accuracy of estimates than the proportion of the population sampled, particularly in the case of dictionary sampling where the sample size is not close to the size of the total dictionary. If we want to be sure that the true value of an individual learner's score on a test using the sample lies close to the learner's observed score, we need to have a large enough sample. Table 1 uses calculations based on the normal approximation to a binomial distribution to show the effect of sample size on the maximum range of true values on either side of the observed score.

Table 1 shows that if the sample size was 100 words, then we are reasonably confident that the true value of an individual learner's observed score would lie anywhere within a 16% range. Thus if the learner's observed score on the test was 50 out of 100 (50%), we could be 90% sure that the true value of his or her score lay between

Table 1 Confidence intervals at the 90% level for different sample sizes for simple random sampling

Number of items in the sample	Maximum confidence interval*
100	±8%
150	±7%
200	±6%
300	±5%
400	±4%
600	±3%
1200	±2%

Note: * i.e., we can be 90% sure that the true value of the score lies within ± this percentage of the observed score.

42 (42%) and 58 (58%) out of 100 (i.e. a range of ±8%). However, if a sample of 600 items was used and a learner scored 300 out of 600 (50%), we could be 90% sure that the true value of the learner's score lay somewhere between 282 (47%) and 318 (53%) (i.e. a range of ±3%). Note that although the range in number of items is larger, the percentage is smaller, and the percentage is used when estimating from a sample to the whole dictionary. According to the values in Table 1, to narrow the range of the true value usefully further, the sample size would have to be increased to 1 200 items. This is too large for a manageable test of vocabulary size.

The size of the maximum confidence interval has an important effect on the usefulness of a vocabulary-size test. *Collins English dictionary*, a medium-sized dictionary, contains 31000 base words. A one-in-50 sample of this dictionary would contain 622 items. A test using this number of items would have a confidence interval of ±3%. This is a 6% range which represents 1866 items in the 31100 pool. The best information at present available suggests that the vocabulary size of native speakers of English grows by about 1000 base words a year (Goulden, Nation, and Read, 1990; D'Anna and Zechmeister, forthcoming). A test containing 622 items could not be used to measure confidently yearly and possibly two-yearly increases in an individual's vocabulary size because the amount of increase would be within the range of the confidence interval.

In order to test this number of items within a feasible length of time it would be necessary to use test items that could be responded to very quickly, to computerize the test perhaps using coarse and fine measures (Meara and Jones, 1987), and to order the test words according to their frequency of occurrence in the language so that sections of them could be assumed to be known and need not be tested.

With group measures of vocabulary size, different calculations involving ANOVA, possibly nonparametric, are required and the word sample sizes for each individual could be much smaller. The word sample size, the subject sample size (people sitting the test), and the range of variability in the scores would have to interact to produce a maximum confidence interval of ±0.5% for the proportion of the dictionary known. This is because a dictionary needs to contain at least 30000 words (step 1). As probable yearly increases could range from 300 to 1000 words, a one-in-50 sample (600 items) with a confidence level of ±0.5% would be needed to measure yearly increases. This means that a learner's score on the test would probably increase by six to 20 items (representing 300 to 1000 words in the dictionary) and six out of 600 is 1% or ±0.5%. In group measures it is better to have a smaller word-sample size and a large subject-sample size than a larger word-sample size and fewer subjects. The greater the range of variability in the test scores, the greater the number of subjects needed to sit the test. Sample sizes cannot be estimated without knowing the variability of vocabulary size between individuals within groups. To be balanced against this, however, is the need to have a large enough sample size to provide a ratio of sample to population that is sensitive enough to measure change.

6 The sampling should be checked for the reliability of the application of the criteria for exclusion and inclusion of items

There are several ways of carrying out such checks, particularly to see if the criteria for deciding what is to be included as words are consistently applied. For example, the sample can be done in sections and the figures for each section compared, or more than one person can do part of the sampling and an inter-rater reliability check can be carried out. If the decision being made is simply whether an item is a base word or not, then the accuracy of the classification could be checked at the 0.01 level of significance by the other rater working on a total of 14 items (seven base words and seven nonbase words). A 0.85 agreement is tolerable for data with two response alternatives (Rosenthal, 1987: 64–67). If there are more than two categories, for example base word, derived word, proper noun, compound word, then fewer items would need to be checked in each category, although the total checked would rise compared to the total for two category checks. With more categories the tolerable level of agreement could be a little less than 0.85. This internal check is not a substitute for an external check using a frequency list.

7 The sample should be checked against a frequency list to make sure that there is no bias in the sampling towards high-frequency items

If the sampling is properly carried out according to the procedures described in this article, then when the sample is compared to the words in a frequency count, such as Thorndike and Lorge (1944), Carroll, Davies and Richman (1971), Francis and Kucera (1982), there should be proportionally the appropriate number of words at each frequency level in the sample. For example, if each word in the sample represents 100 words in the dictionary then there should be 10 words in the sample in the 1000 most frequent words in the frequency list (10×100). Similarly, there should be 10 words in the sample from the second 1000 most frequent words, and so on.

This procedure of checking was used by Thorndike (1924), Williams (1932) and Lorge and Chall (1963) and in all cases revealed bias in the studies they checked. Goulden, Nation and Read (1990) did a similar check and found it necessary to recalculate the number of words at the various levels in the Thorndike and Lorge (1944) frequency count to match the definition of a word family used in the sampling from the dictionary. Thorndike and Lorge defined a word family as including the base form and the base form plus the inflectional suffixes *-s*, *-ed*, *-ing*, *-er*, *-est*, *-ly*. Goulden, Nation and Read included many more items in their definition of what was a word family. So, derived items that were separate entries in the Thorndike and Lorge count, such as *lengthen* and *lengthwise*, were included in the same word family and excluded from the Goulden, Nation and Read sample. When this same way of making a word family is applied to the Thorndike and Lorge count, the 30000 items reduce to 13 900 items with 6100 in the first 10000, 3600 in the second 10000, and 4200 in the third 10000. Once these necessary recalculations were made, Goulden, Nation and Read (1990) and Nation and Ellis (in preparation) found that their samples were not biased towards high-frequency items.

8 In the written report of the study, describe clearly and explicitly how each of the previous seven procedures was followed in sufficient detail to allow replication of any or all of the procedures

Using the eight procedures described above, it is possible to review studies of vocabulary size and see where faulty methodology led to faulty estimation. Such reviews (Thorndike, 1924; Lorge and Chall, 1963) agree with Goulden, Nation and Read's (1990) estimate that, as a very rough rule of thumb, for people up to the age of 16 years

we can multiply their age in years by 1000 base words or less per year to get an indication of probable vocabulary size. More importantly, however, the procedures should be used to carry out methodologically sound studies that will provide information that can be relied upon in order to explore the many roles of vocabulary knowledge in intellectual activity.

This article has focused on the methodology of sampling. This is just one of several issues in the measurement of vocabulary size. Others, also noted by Thorndike (1924), are 'What do we mean when we say someone knows a word?', 'What test item types should we use to measure this knowledge?' and 'What is the significance of vocabulary size?' These issues need to be considered in systematic and explicit ways drawing on the results of previous studies. In this way we can make progress in this difficult and important area of research.

Acknowledgement

I am grateful for the help I received from Stephen Haslett of the Institute of Statistics and Operations Research and James Dickie of the English Language Institute in the preparation of this paper.

IV References

Anderson, R.C. and **Freebody, P.** 1983: Reading comprehension and the assessment and acquisition of word knowledge. *Advances in Reading/Language Research* 2, 231–56.

Bock, C. 1948: Prefixes and suffixes. *Classical Journal* 44, 132–33.

Brandenburg, G.C. 1918: Psychological aspects of language. *Journal of Educational Psychology* 9, 313–32.

Carroll, J.B., Davies, P. and **Richman, B.** 1971: *The American heritage word frequency book.* New York: American Heritage Publishing Co.

D'Anna, C.A. and **Zechmeister, E.B.** Forthcoming: Toward a meaningful definition of vocabulary size. *JRB: A Journal of Literacy.*

Diller, K.C. 1978: *The language teaching controversy.* Rowley, MA: Newbury House.

Dolby, J.L. and **Resnikoff, H.L.** 1967: *The English word speculum.* The Hague: Mouton & Co.

Dupuy, H.J. 1974: *The rationale, development and standardization of a basic word vocabulary test.* Washington, DC: US Government Printing Office.

Francis, W.N. and **Kucera, H.** 1982: *Frequency analysis of English usage.* Boston: Houghton Mifflin Co.

Gerlach, F.M. 1917: Vocabulary studies. *Studies in education and psychology* 1. Colorado College.

Goulden, R., Nation, P. and **Read, J.** 1990: How large can a receptive vocabulary be? *Applied Linguistics* 11, 341–63.

Kirkpatrick, E.A. 1891: Number of words in an ordinary vocabulary. *Science* 18 (446), 107–108.

Lorge, I. and **Chall, J.** 1963: Estimating the size of vocabularies of children and adults: an analysis of methodological issues. *Journal of Experimental Education* 32, 147–57.

Meara, P. and **Jones, G.** 1987: Tests of vocabulary size in English as a foreign language. *Polyglot* 8, fiche 1.

Nagy, W.E. and **Anderson, R.C.** 1984: How many words are there in printed school English? *Reading Research Quarterly* 19, 304–30.

Nation, P. and **Ellis, B.** In preparation: vocabulary growth.

Rosenthal, R. 1987: *Judgement studies: design, analysis, and meta-analysis.* Cambridge: Cambridge University Press.

Seashore, R.H. and **Eckerson, L.D.** 1940: The measurement of individual differences in general English vocabularies. *Journal of Educational Psychology* 31, 14–38.

Stauffer, R.G. 1942: A study of prefixes in the Thorndike list to establish a list of prefixes that should be taught in elementary school. *Journal of Educational Research* 35, 453–58.

Thorndike, E.L. 1924: The vocabularies of school pupils. In Carelton Bell, J., editor, *Contributions to education*, New York: World Book Co., 69–76.

Thorndike, E.L. and **Lorge, I.** 1944: *The teacher's word book of 30 000 words*. New York: Teachers' College, Columbia University.

Tyler, A. and **Nagy, W.** 1989: The acquisition of English derivational morphology. *Journal of Memory and Language* 28, 649–67.

White, T.G., Power, M.A. and **White, S.** 1989: Morphological analysis: implications for teaching and understanding vocabulary growth. *Reading Research Quarterly* 24, 283–304.

Williams, H.M. 1932: Some problems of sampling in vocabulary tests. *Journal of Experimental Education* 1, 131–33.