

TOEIC & TOEFL Vocabulary Secrets Revealed

July 2013

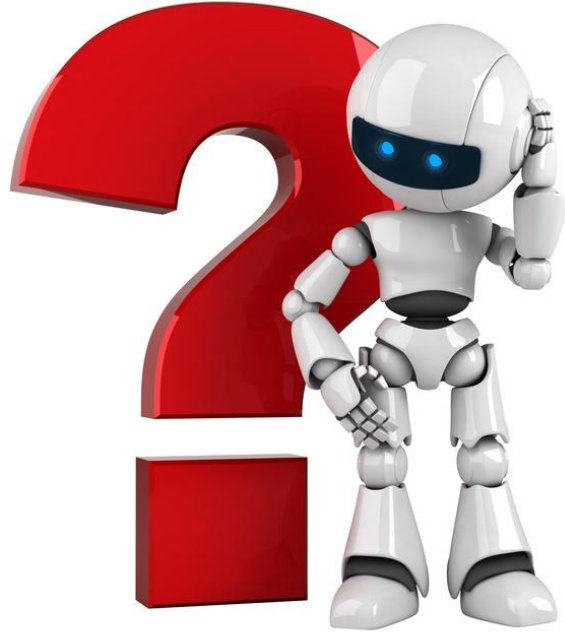
Presenter: Guy Cihi



Lexica R&D
2-7-8 Shibuya 5F
Shibuya-ku, Tokyo 150-0002
info@lexica.co.jp



Hello, I'm Roby!



Who is Lexxica?



Lexica offers two consumer programs.



Learn spoken English. For kids of all ages!



WordEngine™
ワードエンジン

High Speed Vocabulary Learning System

Learn more than 100
new words per day!



lexica™

Founders and important contributors



Guy Cih CEO, Co-Founder



Dr. Charles Browne
Co-Founder



Dr. Brent Culligan
Senior Scientist



Dr. Paul Nation
Unpaid Advisor

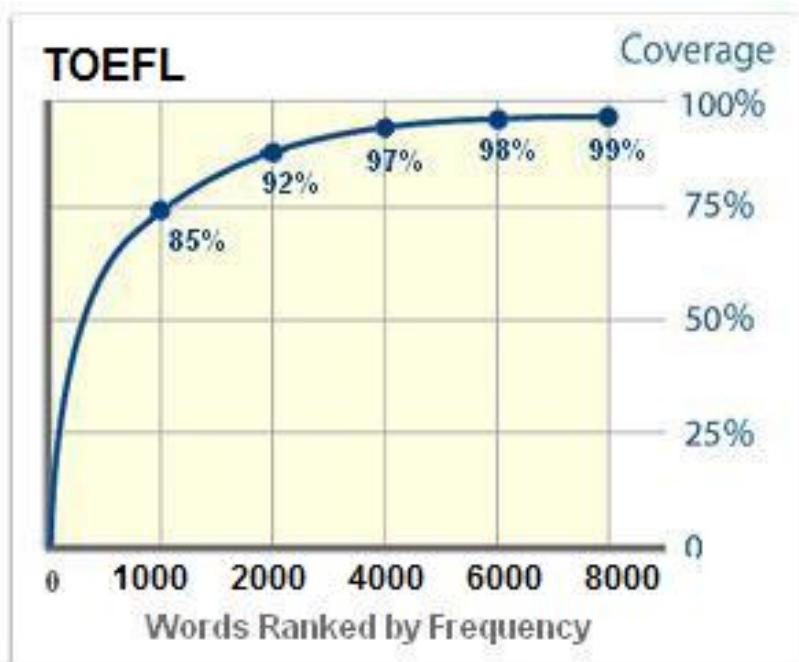


Della Summers
Senior Editor



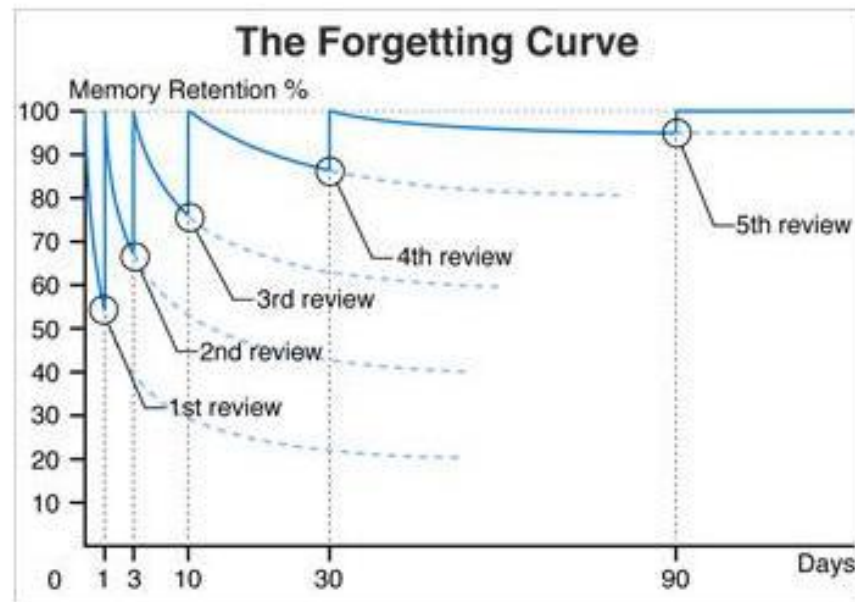
Prof. Bruce Rogers
Senior Testing Advisor

Two foundational concepts



Coverage

Certain words occur more frequently than others. The words that occur most often provide a tremendous advantage to learners. Corpus analysis shows which words occur most frequently in each subject. We teach up to 99% coverage of each subject.



The Forgetting Curve. Dr. Hermann Ebbinghaus

Memorization

There is proven science regarding how to quickly convert short term memory into long term memory. The high frequency words we teach are repeated at the specific time intervals proven to efficiently inculcate long term memory.

lexica

We do our own corpus analysis work

We study exactly which words are required to master each subject area.

All General English

13,384 words

Business English

8,742 words



College Entrance

5,435 words

High School

3,552 words

Elementary

2,000 basic words

TOEFL

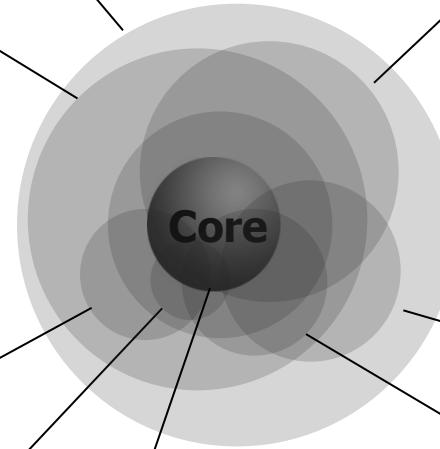
7,501 words

TOEIC

6,480 words

IELTS

5,870 words



TOEIC Corpus Analysis

1,250,000 total words

14,652 different words

**6,480 different words constitute
99% of all occurrences**

**982 different words constitute
90% of all occurrences.**

**These 982 are the absolutely essential
Super-High-Frequency words of TOEIC**



TOEFL Corpus Analysis

1,250,000 total words

16,736 different words

**7,501 different words constitute
99% of all occurrences**

**1,513 different words constitute
90% of all occurrences.**

**These 1,513 are the absolutely essential
Super-High-Frequency words of TOEFL**

My question to you



Does increased vocabulary size improve listening, speaking, reading, and writing?



The majority consensus is **YES**, increased vocabulary size does improve listening, speaking, reading, and writing!



Word Engine

The program begins with a 5 minute needs assessment called V-Check



The screenshot shows a software interface for a needs assessment. At the top left is the logo "V-Check®". To its right is a progress bar with "Start" at the beginning and "End" at the end. In the center is a large, light-yellow rectangular box containing the word "angel" in a dark blue font. Below this box, the text "Do you know the meaning of this word?" is displayed. At the bottom are two blue buttons with white text: "yes" on the left and "no" on the right.

V-Check®

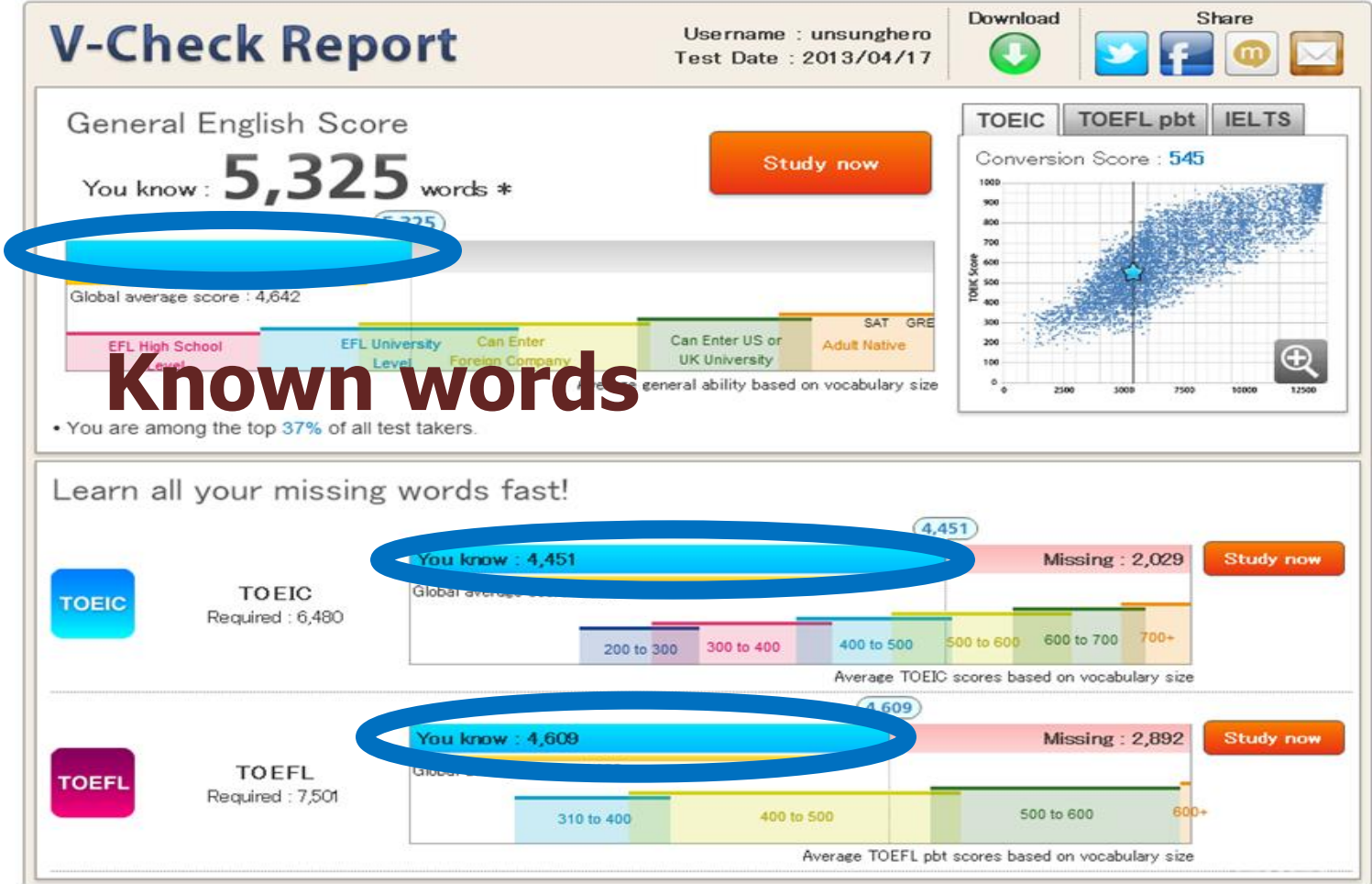
Start End

angel

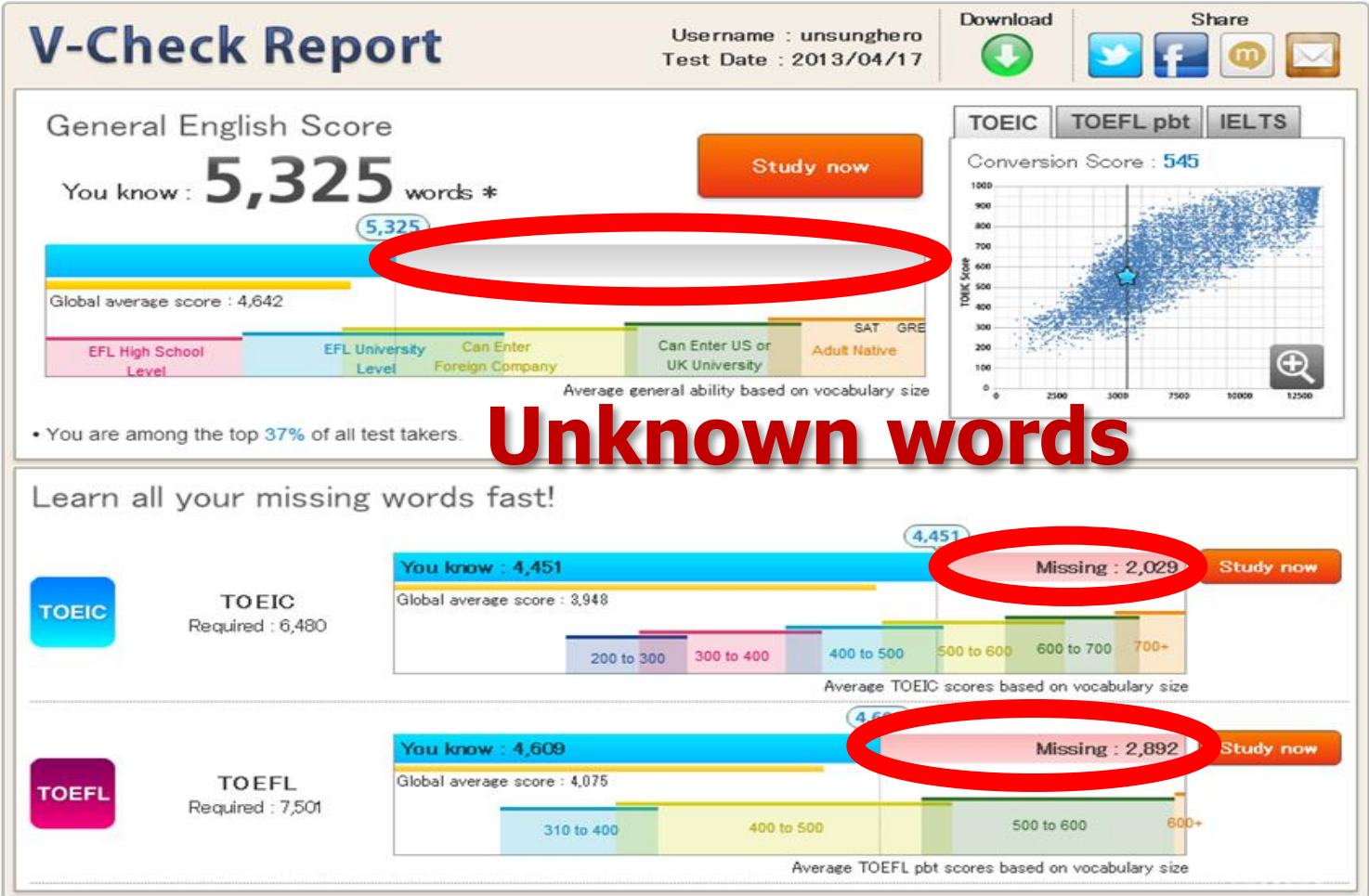
Do you know the meaning of this word?

yes no

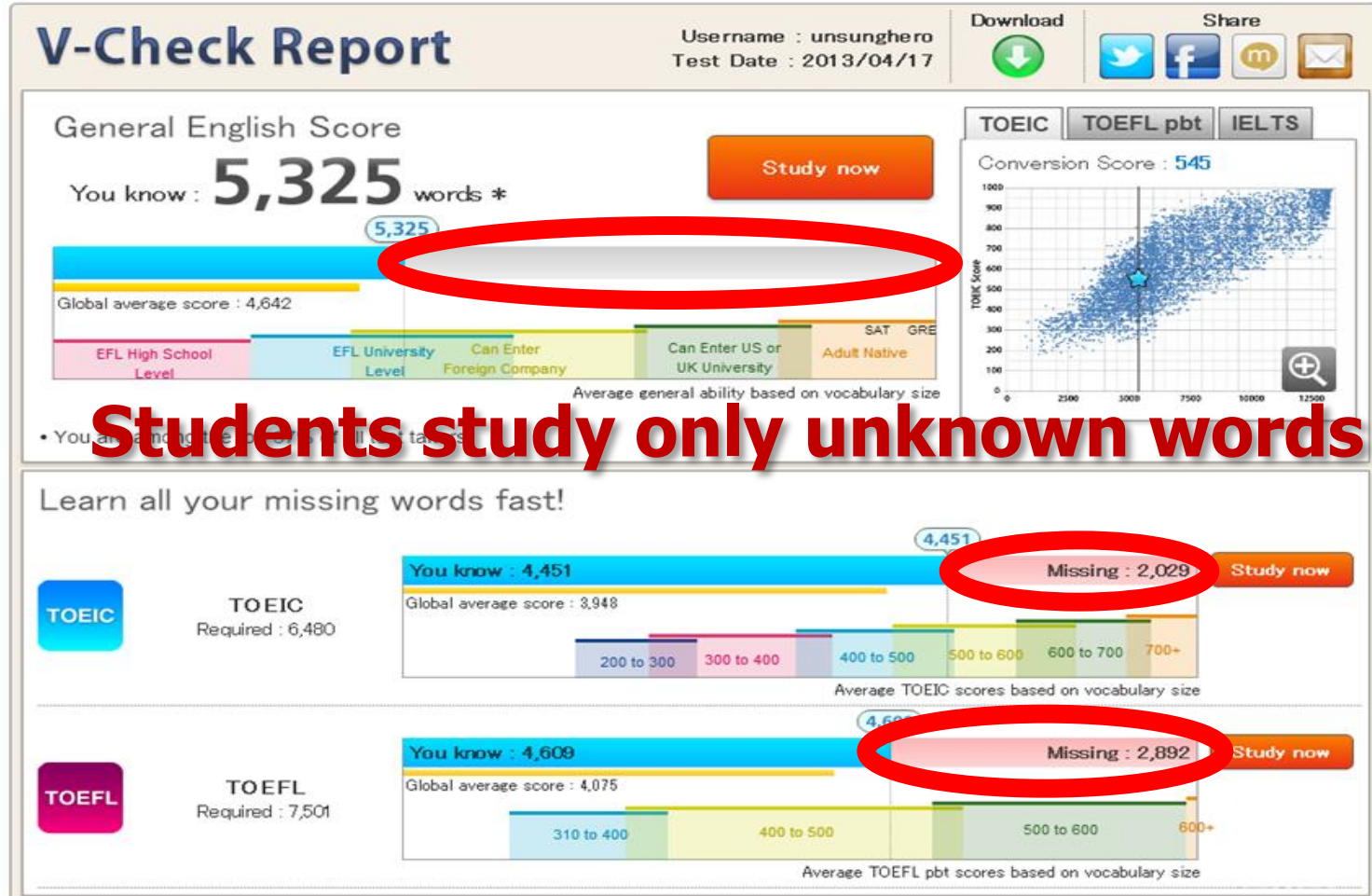
V-Check identifies each students known and unknown words.



V-Check identifies each students known and unknown words.



V-Check identifies each students known and unknown words.





ck I
English
5,
ore : 4,64
ol
the top 2
our n

特許証
(CERTIFICATE OF PATENT)

特許第4908495号
(PATENT NUMBER)

発明の名称 (TITLE OF THE INVENTION)	意味論的知識の評価と指導と習得とに関するシステムおよび方法
特許権者 (PATENTEE)	中華人民共和国 香港 セントラル ハリウッド ロード 32 キンウィック センター 1601 国籍 中華人民共和国 エーアイ リミテッド
発明者 (INVENTOR)	ガイ シヒ チャールズ ブラウン ブレント カリガン
出願番号	特願2008-504874

その他別紙記載



oad
load
IC
TOEFL pb
version Score : 5

One-parameter Logistic Model

Using the Rasch model, we can calculate the probability of an examinee answering an item correctly with the following formula:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability of a randomly chosen examinee with ability θ answering item i correctly.

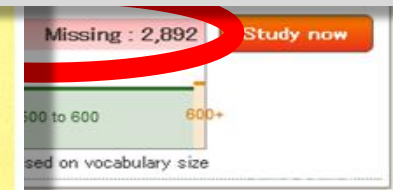
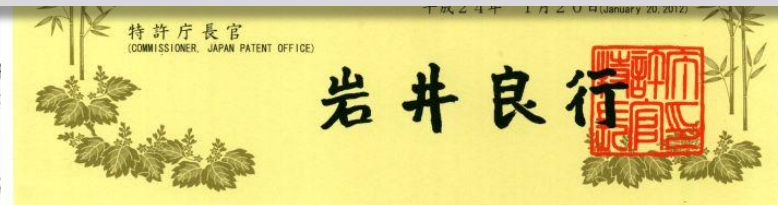
e is the base of natural logarithms (2.718)

θ is the person ability measured in logits

b_i is the difficulty parameter of the item measured in logits

ing words

Lexica holds international patents for the process of identifying which specific words are known and unknown.



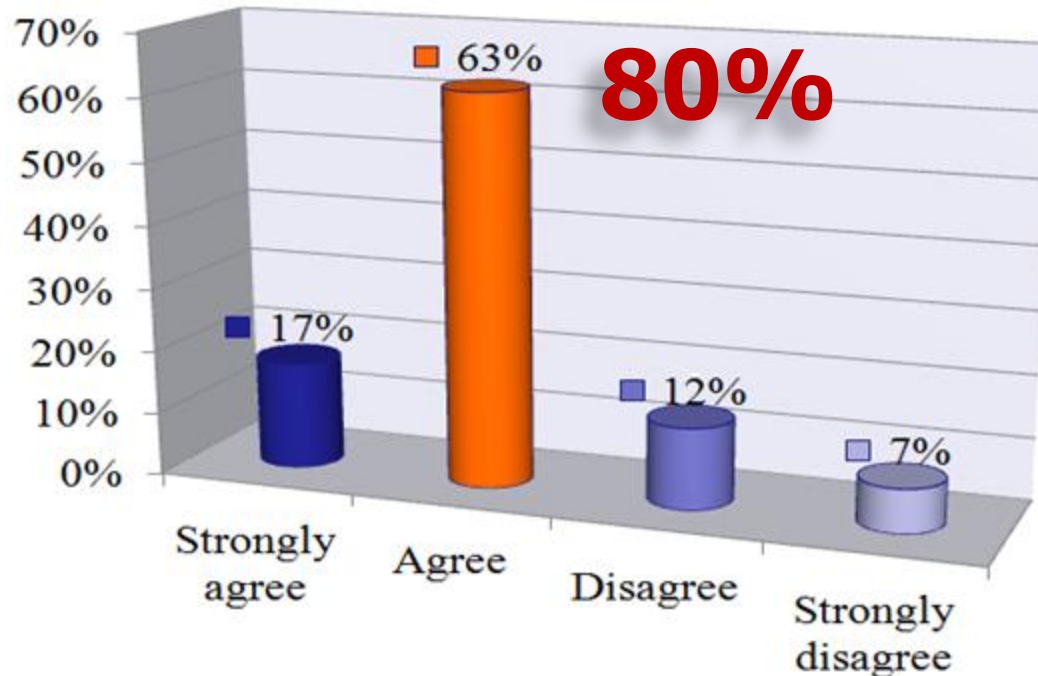


SightWords - Focus on Visual Speed



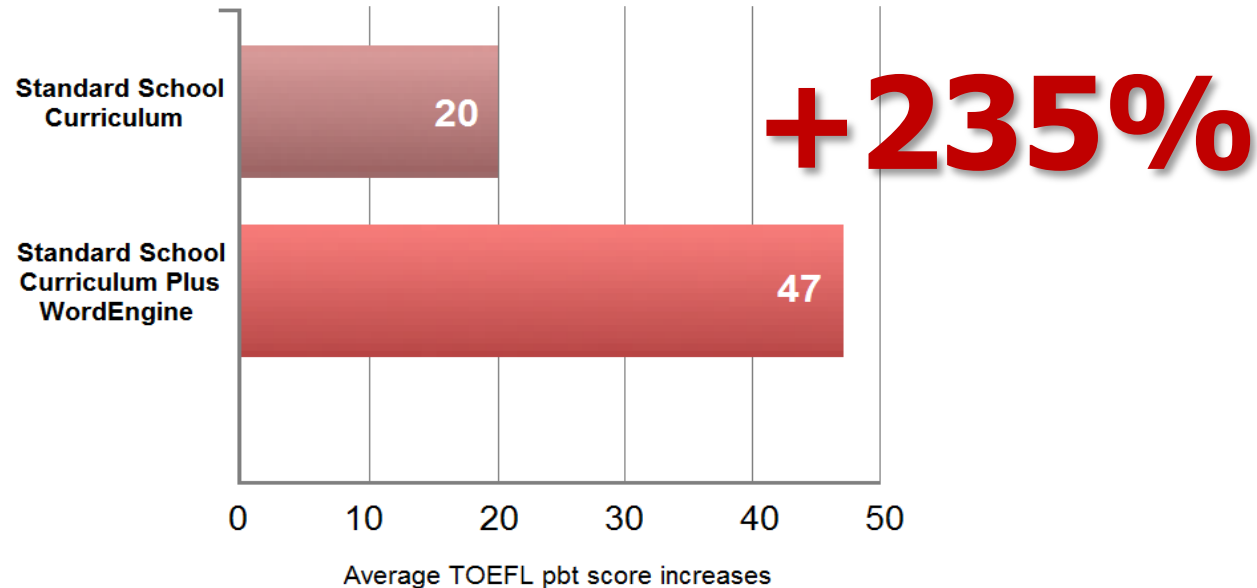
The average study pace is 360 words per hour

WordEngine is a fast way to learn new TOEIC vocabulary



Actual TOEFL pbt score improvements

for students who scored above 400 on their first TOEFL



Source: Dr. David Coulson, University of Nigata Prefecture 2011

Main Presentation

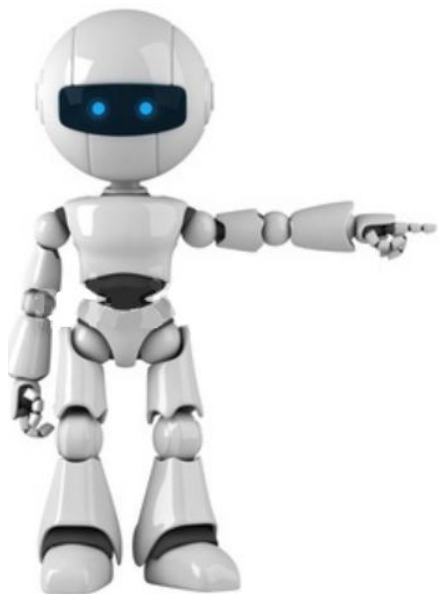
TOEIC and TOEFL Vocabulary Secrets



Secret #1

TOEIC and TOEFL are Item Response Theory Proficiency Tests – not English ability diagnostic tests. **These tests are not designed to provide meaningful advice for improving English ability.**

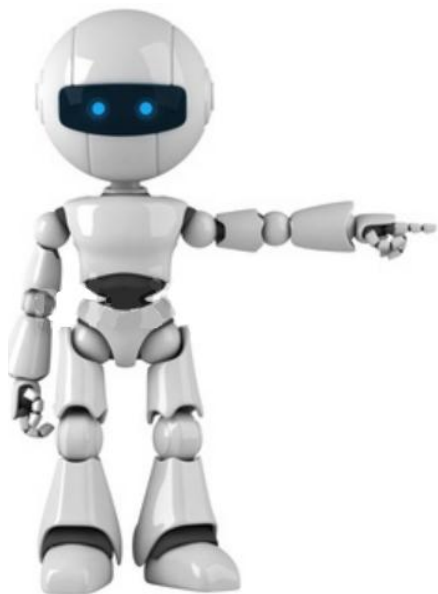
Students are scored based on their correct responses to questions having known difficulty metrics. The difficulty metrics are established through statistical analysis of all prior uses of each question.



Secret #2

Without a full range of questions from easy to difficult, Education Testing Service “ETS,” would be unable to maintain its bell-curve and generate ‘reliable’ scores.

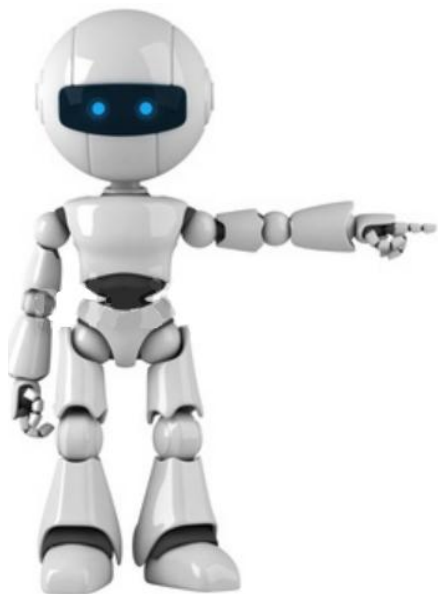
It is impossible to write statistically difficult questions. Only field testing can identify the difficulty of questions.



Secret #3

95% of test questions are recycled.
5% are new questions that are in the process
of being measured for difficulty.

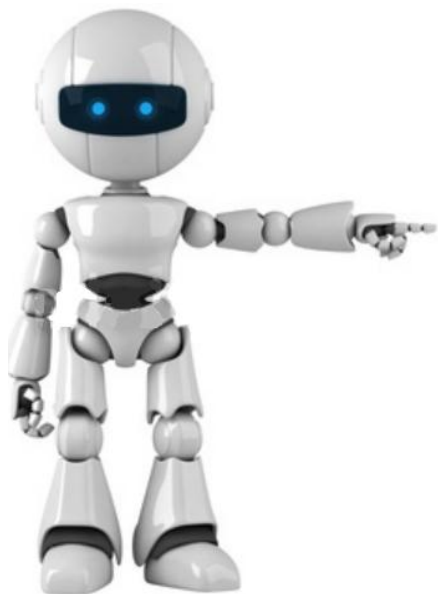
**The 95% recycling requirement means
that vocabulary on the tests can be
accurately predicted.**



Secret #4

ETS has never, and likely will never issue a vocabulary guide for any of its major tests including: TOEIC, TOEFL, SAT and GRE.

Why?



Secret #4

Because using difficult words, and irregular definitions, are great ways to create a wide variety of questions at all levels of difficulty.

Publishing an official vocabulary guide would both expose a scoring system vulnerability and defeat the purpose of their tests which is to measure familiarity and proficiency with authentic English.

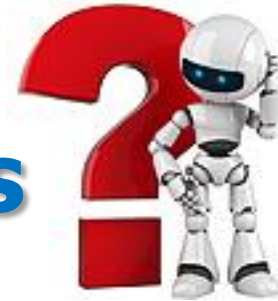


TOEIC, TOEFL (and IELTS) versus General English

**1/3 of the words in all parts of TOEIC
and TOEFL are not common, high
frequency words in General English.**

(1/4 of the words in IELTS.)

What kinds of words



Top 2000 high frequency words of TOEIC and General English

TOEIC

ability

able

aboard

about

above

abroad

absence

absent

absolutely

abstract

accept

General

ability

able

about

above

abroad

absence

absolute

absolutely

absorb

abuse

academic

accept

TOEIC

gain

gallery

gallon

game

garage

garbage

garden

gardener

gas

gasoline

gate

gather

gender

general

General

gain

gall

game

gap

garage

garden

gas

gate


gather

gaze

gear

gene

general

 Frequent only in the TOEIC corpus.

 Frequent only in the General corpus.

Our general corpus contains 850 million words from all genres.

What does this mean?



EFL students can't learn the words they need because they aren't in their study and reading materials.

Because their study materials are simplified.

We used to say:

Education Testing Service (ETS) purposefully uses difficult words and seldom used meanings of common words because otherwise their scoring system fails.

Then we talked to ETS authors

Now we say:

Education Testing Service (ETS) ~~purposefully~~ uses difficult words and seldom used meanings of common words because otherwise their scoring system fails.

To create the test questions:

Authors are told to search through authentic materials to find texts and dialogs to adapt for the different types of test questions.

(They are also told to change details such as names and locations.)

To evaluate new test questions:

When finished, the authors don't know how difficult their new questions are.

The only way to find out is for ETS to put them into actual tests alongside questions for which they do know the difficulty.

The science of question difficulty:

A one-parameter Logistic Model is used to calculate the difficulty of each question.

Using the Rasch model, we can calculate the probability of an examinee answering an item correctly with the following formula:

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}}$$

where $P_i(\theta)$ is the probability of a randomly chosen examinee with ability θ answering item i correctly.

e is the base of natural logarithms (2.718)

θ is the person ability measured in logits

b_i is the difficulty parameter of the item measured in logits

Testing the test questions:

On every TOEIC and TOEFL test 5% of the questions are new questions that have no affect on scoring.

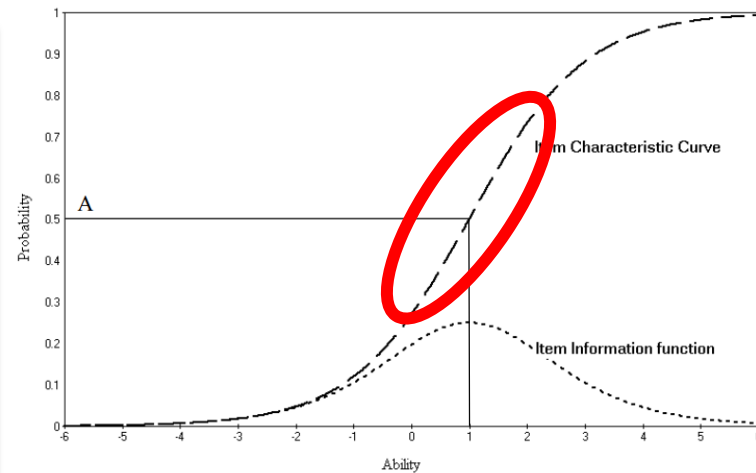
95% are recycled questions that have known and reliable difficulties that can be used for scoring.

95% regularly recycled questions make these tests ideal for Lexxica's predictive corpus modeling

Item Response Theory

An IRT test only works when it has many items at all levels of difficulty.

Figure 1. Characteristic Curve and Information Function of the Item

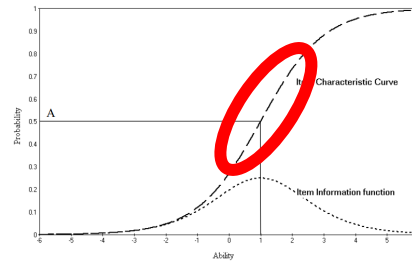


Difficulty/Ability/Score

ETS's Primary Concern

ETS's primary concern is that their scores accurately reflect each respondent's **relative** familiarity and proficiency with **authentic** English.

Figure 1. Characteristic Curve and Information Function of the Item

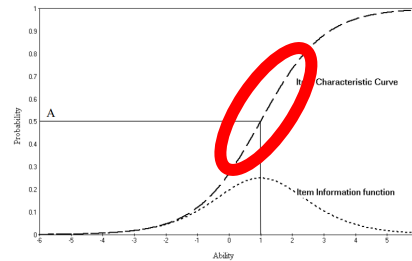


From corpus analysis we know:

1/3 of the words on TOEIC and TOEFL tests are low frequency 'authentic' vocabulary words.

Vocabulary is the main reason one test question is more or less difficult than another.

Figure 1. Characteristic Curve and Information Function of the Item



Note that many of the 1/3 low frequency words have multiple meanings

TOEIC

ability

able

aboard

about

above

abroad

absence

absent

absolutely

abstract

accept

General

ability

able

about

above

abroad

absence

absolute

absolutely

absorb

abuse

academic

accept

TOEIC

gain

gallery

gallon

game

garage

garbage

garden

gardener

gas

gasoline

gate

gather

gender

general

General

gain

gall

game

gap

garage

garden

gas

gate


gather

gaze

gear

gene

general

 Frequent only in the TOEIC corpus.

 Frequent only in the General corpus.

Our general corpus contains 850 million words from all genres.

Typical low frequency definition: **crack**

A line along which something has split without breaking into separate parts: “a crack in the surface”

An illegal street drug: “He was arrested for crack.”

ETS used this:

To open something after making a concerted effort: “to crack a safe”

Very good, esp. at a specified activity: “He’s a crack shot.”

Why use low frequency definitions?

ETS's primary concern is that scores accurately reflect each respondent's relative familiarity and proficiency with authentic English.

ETS's advice for scoring higher on TOEIC and TOEFL is to read authentic texts.

(Graded readers won't help because they're simplified)

How much authentic text?

According to research by Rob Waring, they'll need to read 6,250 hours of authentic text in order to meet the test words often enough to memorize them.

A statistical analysis of the number of English words you need to meet (at given recurrence rates) to 'learn' that number of words

A Word rank	B Percentage of general English that this word covers	C (= 100 / B)	D (= x times C)				E (= D / Book length)				
			Volume of text you need to read to meet the words at these recurrence rates				Number of books to cover this volume given these recurrence rates				
			5 times	10 times	20 times	50 times	Book length	5 times	10 times	20 times	50 times
1 st most frequent (<i>the</i>)	5.83898%	17 (1)	86	171	343	856	4,500	0.0	0.0	0.1	0.2
2 nd most frequent (<i>be</i>)	5.12332%	20	98	195	390	976	4,500	0.0	0.0	0.1	0.2
25 th (<i>as</i>)	0.44382%	225	1,127	2,253	4,506	11,266	4,500	0.3	0.5	1.0	2.5
50 th (<i>like</i>)	0.24109%	415	2,074	4,148	8,296	20,739	4,500	0.5	0.9	1.8	4.6
100 th (<i>hear</i>)	0.10505%	952	4,759	9,519	19,038	47,595	4,500	1.1	2.1	4.2	10.6
500 th (<i>present</i>)	0.02477%	4,037	20,183	40,366	80,732 (4)	201,829	4,500	4.5	9.0	17.9	44.9
1000 th (<i>blood</i>)	0.01172%	8,533 (3)	42,665	85,329	170,658	426,645	10,000	4.3	8.5	17.1	42.7
1500 th (<i>intent</i>)	0.00677%	14,773	73,864	147,727	295,455	738,636	15,000	4.9	9.8	19.7	49.2
2000 th (<i>stumble</i>)	0.00432% (2)	23,103	115,625	231,250	462,500	1,156,250	20,000	5.8	11.6	23.1	57.8
3000 th (<i>sergeant</i>)	0.00211%	47,343	236,713	473,425	946,850	2,367,126	30,000	7.9	15.8	31.6	78.9
5000 th (<i>satellite</i>)	0.00076%	132,143	660,714	1,321,429	2,642,857	6,607,143	80,000	8.3	16.5	33.0	82.6
10,000 th (<i>relativity</i>)	0.00016%	632,000	3,164,474	6,328,947	12,657,895	31,644,733	80,000	39.6	79.1 (5)	158.2	395.6

Examples: (1) The most frequent word in English (*the*) covers 5.839% of any general English text (i.e. it occurs once in every 17 words).
 (2) The 2000th most frequent word in English covers 0.00432% of any general English text (and occurs once every 23,103 words).
 (3) To meet all the 1000 most frequent words in English once, you'd need to read 8,533 words.
 (4) To meet all the 500 most frequent words in English 20 times, you'd need to read 80,732 words.
 (5) To meet all the 10,000 most frequent words in English 10 times, you'd need to read 79.1 books that are 80,000 words long.

© R. Waring 2007

If you read at 60 words per minute...

You must read for **2 hours every day**
for... **8.5 years.**

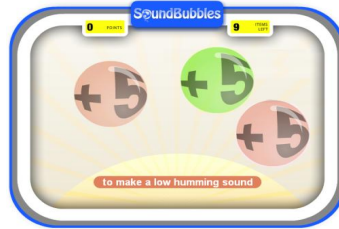
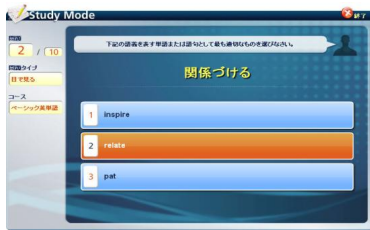
There is a faster way!

The fastest way is:

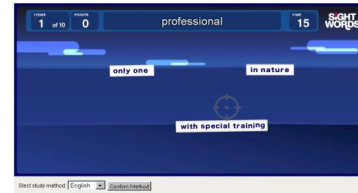
WordEngine™
ワードエンジン

High Speed Vocabulary Learning System

www.wordengine.jp



SightWords - Focus on Visual Speed



Proven capable of teaching one new word per minute!

For acquiring low frequency, domain specific vocabulary, WordEngine is at least 400 times faster than reading authentic texts.



www.wordengine.jp