Running E-Z Reader Simulations

The program for running simulations using the E-Z Reader model (ver. 10) was written in Java, version 1.6. The executable (*.jar*) version of the program is available at: www.pitt.edu/~reichle/ezreader.html. The source code (i.e., *.java* classes) is also available from me (Erik Reichle, at: reichle@pitt.edu) upon request. The first part of these instructions describes how to use the program file to run simulations. The second part describes how to set up your own sentence/text files.

1. Running Simulations

You will need three files to run E-Z Reader 10 simulations: (1) the program file containing the actual model (*E-Z_Reader_10.jar*), (2) a file containing the sentences/text that will be used in the simulation (e.g., *SRC98Corpus.txt*), and (3) a file used to identify specific target words of interest (e.g., *SRC98Targets.txt*). To run a simulation, first download all three files to a common location on your computer (e.g., its desktop or common folder) and then double-click on the program file. This should open a graphic-user interface (GUI) with buttons and text fields that can be selected or modified for running different types of simulations. Here is a brief explanation of the GUI, the different types of simulations that can be run, and of how to create your own sentence/text files and target-word files:

The only information that must be entered into the GUI before you can start running a simulation is the following:

(1.) *Corpus File Name* – Enter the name of the file containing the sentences/text that will be used in the simulation. The file that is provided on my website (*SRC98Corpus.txt*) contains the sentences that were first used by Schilling, Rayner, and Chumbley (1998) in their eye-movement experiment, and that have subsequently been used by me and my colleagues to evaluate different versions of the E-Z Reader model.

(2.) *Target Word File Name* – Enter the name of the file that is used to identify the target words that are of interest in a simulation. This file consists of a single column of integers, with each integer indicating the ordinal position of one target word per sentence. For example, in *SRC98Targets.txt*, the file begins with "5, 5, 3, 9…" arranged in a vertical column. These numbers identify the target words used by Schilling et al. (1998) in their eye-movement experiment and that are contained in the corpus file *SRC98Corpus.txt*. These target words are "apartment," "president," "students," and "baseball," respectively. Please note that—as per Java conventions—both the sentences and words are numbered starting with zero, so that word #5 in the first sentence is actually the sixth word in the sentence. Also note that the program will not run without a target word file (i.e., the field labeled "Target Word File Name:" cannot be left empty). If you are not interested in specific target words, then simply create and enter the name of a dummy file (e.g., a file containing a column of zero, one per sentence) to run simulations.

(3.) # *Subjects* – Enter the number of subjects (1-10,000) that will be used in completing the simulation.

(4.) *RUN* – Press this button to start the simulation. The progress bar at the bottom of the GUI will indicate how much of the simulation has completed. The length of time that is needed to complete a simulation will obviously depend upon the speed of your computer, and on both the number of statistical subjects and the length of sentences/text being used in the simulation. With my 2.33 GHz MacBook Pro, a simulation using the Schilling et al. (1998) sentences and 1,000 statistical subjects takes approximately 9 seconds to complete.

All of the other buttons and text fields are set to default values, but can be modified as necessary. The function of the text field labeled *Output File Name* is fairly obvious; the results of any simulation are written to this file (default name: *Simulation Results.txt*), which will appear in the same location as the program. The functions of the text fields in the box labeled *Parameter Values* is also pretty obvious; you can change the default parameter values of the E-Z Reader model by entering new values. For a description of the parameters and their interpretation, see Reichle, Warren, & McConnell (2009). A few notes on parameter values:

i. The parameter that controls the variability of the gamma distributions, $\sigma_\gamma = 20$, is set equal to a value that generates gamma distributions having standard deviations that are equal to 0.22 of their means. For more information about the gamma distribution function that is used in the E-Z Reader program, see Press, Teukolsky, Vetterling, & Flannery (1992). *Numerical Recipes in C: The Art of Scientific Computing*. New York: Cambridge University Press.

ii. To disable post-lexical processing,

Finally, the buttons in the box labeled *Simulation Output?* allow different types of simulation results to be written to the output file after a simulation has completed. The different types of output are as follows:

(1.) *Word IVs* – Selecting this button will cause all of the independent variables that are contained in the sentence/text file and those that are calculated by the model program prior to executing a simulation (e.g., each word's optimal viewing position) to be written to the output file. It's a good idea to run a simulation with this button selected prior to completing any other simulations to ensure that the sentence/text file has been formatted correctly. (It is also a good idea to use a very small number of statistical subjects to avoid creating a really big text file.) Here is an example of the simulation output and an explanation of what it means:

Sentence: 0

#	Word	Freq	lnFreq	Class	Pred	Length	Char0	Char1	CharN	OVP
0	Margie	1	0.00	1	0.0	6.0	0.0	1.0	7.0	4.0
1	moved	181	5.20	3	0.0	5.0	7.0	8.0	13.0	10.5
2	into	1789	7.49	4	0.2	4.0	13.0	14.0	18.0	16.0
3	her	3036	8.02	4	0.25	3.0	18.0	19.0	22.0	20.5

4	new	1635	7.40	4	0.65	3.0	22.0	23.0	26.0	24.5
5	apartment	81	4.39	2	0.75	9.0	26.0	27.0	36.0	31.5
6	at	5372	8.59	4	0.0	2.0	36.0	37.0	39.0	38.0
7	the	69974	11.16	5	0.6	3.0	39.0	40.0	43.0	41.5
8	end	409	6.01	3	0.1	3.0	43.0	44.0	47.0	45.5
9	of	36414	10.50	5	0.95	2.0	47.0	48.0	50.0	49.0
10	the	69974	11.16	5	1.0	3.0	50.0	51.0	54.0	52.5
11	summer.	134	4.90	3	0.1	7.0	54.0	55.0	62.0	58.5

Sentence: 1

#	Word	Freq	lnFreq	Class	Pred	Length	Char0	Char1	CharN	OVP
0	The	69974	11.16	5	0.0	3.0	0.0	1.0	4.0	2.5
1	principal	92	4.52	2	0.0	9.0	4.0	5.0	14.0	9.5

.
.
.

The above example shows the first sentence (Sentence: 0) and part of the second (Sentence: 1). Following Java conventions, sentences and words are always numbered starting from 0, so that a set of N sentences/words will be numbered from 0 to N-1. The columns from left to right are: (1) word number (#); (2) the actual word (Word); (3) its frequency (Freq); (4) the natural logarithm of its frequency (lnFreq); (5) its frequency class (Class); (6) its predictability (Pred); (7) its length (Length); the cumulative character position of (8) the blank space to the left of the word (Char0); (9) the first character of the word (Char1); (10) the last character (CharN) of the word; and (11) the word's optimal viewing position (OVP).

(2.) *Model States* – Selecting this button will cause the model program to write out all of the internal states that the model progresses through as it simulates the reading of a sentences/text. This type of output is useful for seeing how the model works, and can sometimes be useful for figuring out exactly why the model makes the predictions that it does. Because the output files are very large (each word that is processed might cause the model to go through 6-10 states), it is a good idea to use only a very small number of subjects (e.g., $N = 1$) when running simulations of this type. Here is an example of the simulation output and an explanation of what it means:

```
 Sentence: 0
 Active Processes:
     L1:  duration: 245.37 ms; processing rate: 1.34
 Attention: 0
 Current Fixation:
     word #: 0
     cumulative character #: 3.50
     duration: 0.00 ms
------------------------------
 Sentence: 0
 Active Processes:
     L2:  duration: 62.50 ms
     M1:  duration: 93.13 ms (engage time: 46.57 ms; convert time: 46.57 ms); word target #: 1
 Attention: 0
 Current Fixation:
     word #: 0
     cumulative character #: 3.50
     duration: 245.37 ms
------------------------------
```

Sentence: 0
Active Processes:
    M1: duration: 30.63 ms (engage time: 0.00 ms; convert time: 30.63 ms); word target #: 1
    L1: duration: 223.65 ms; processing rate: 2.74
Attention: 1
Current Fixation:
    word #: 0
    cumulative character #: 3.50
    duration: 307.87 ms

Sentence: 0
Active Processes:
    L1: duration: 193.02 ms; processing rate: 2.74
    M2: duration: 22.40 ms; target word #: 1; intended saccade length: 6.50 characters
Attention: 1
Current Fixation:
    word #: 0
    cumulative character #: 3.50
    duration: 338.50 ms

.
.
.

The above example shows four consecutive model states, each separated by a row of dashes. At the top of each state is the current sentence (the first sentence, or "Sentence: 0", in this example). Below that is a list of all active processes. For V, L_2, A, and I, only the durations (in ms) are shown. For L_1, both the duration and the rate of processing (which varies as a function of retinal eccentricity) are shown. For M_1, three different times are shown: the overall process duration, the time required to engage the oculomotor system (i.e., engage time), the time to convert the (word) spatial target into a distance metric (i.e., convert time). The actual saccade target (i.e., the target word's number) is also shown for M_1. Similarly, for M_2, the duration, the saccade target, and intended saccade length (in character spaces) are shown. The word being attended is shown right below the active processes. Finally, the bottom three lines show the current fixation location (both in terms of the word number being fixated and the cumulative character position) and its duration (in ms). For a detailed discussion of the model states and how state transitions occur in (an earlier version of) E-Z Reader, see Reichle, Pollatsek, Fisher, and Rayner (1998).

(3.) *Fixations* – Selecting this button will result in the output file containing a list of all of the fixations generated by the model—their location in terms of both cumulative character position (i.e., number of characters from the beginning of the sentence) and word number, and their duration (in ms). Again, use a small number of subjects when running this type of simulation. Here is an example of the simulation output and an explanation of what it means:

Sentence: 0; Subject: 0

FixDur	FixLoc	Word#	Fix#
270	4	0	1
211	11	1	2

289	20	3	3
264	29	5	4
224	38	6	5
275	47	9	6

Sentence: 1; Subject: 0

FixDur	FixLoc	Word#	Fix#
282	2	0	1
216	7	1	2
375	17	2	3
310	26	3	4
363	35	5	5
233	44	6	6

Sentence: 2; Subject: 0

FixDur	FixLoc	Word#	Fix#
320	3	0	1
269	10	2	2
421	20	3	3
243	30	5	4
202	36	7	5
273	46	9	6
195	50	10	7

.
.
.

The above output shows the fixations that were generated across three sentences by one statistical subject. The sentence and subject numbers are shown for each sentence. The other information corresponds to individual fixations and is organized into four columns: (1) the duration (in ms) of each fixation (FixDur); (2) its cumulative character location (FixLoc); (3) the word being fixated (Word#); and (4) the fixation number (Fix#).

(4.) *Trace File* – Selecting this button will result in the model generating a trace file that is similar to those that are generated by eye-trackers in experiments involving real human participants. This "trace file" will contain the following information:

Sentence: 13; Subject: 0

Word#	#Fix	#1st	FFD	SFD	GD	TT
0	1	1	316	316	316	316
1	2	1	143	0	143	306
2	0	0	0	0	0	0
3	2	1	92	0	92	347
4	1	1	248	248	248	248
5	0	0	0	0	0	0
6	1	1	182	182	182	182
7	1	1	279	279	279	279
8	1	1	335	335	335	335
9	0	0	0	0	0	0
10	0	0	0	0	0	0
11	0	0	0	0	0	0

.
.
.

The above trace file was generated by allowing the inclusion of inter-word regressions, and by increasing the value of λ to encourage such regressions. The top row indicates the subject and subject numbers. The remaining information is organized into seven columns: (1) a given word's number (Word#); (2) total number of fixations on the word (#Fix); (3) the number of first-pass fixations on the word (#1st); (4) the word's first-fixation duration (FFD); (5) single-fixation duration (SFD); (6) gaze duration (GD); and (7) total viewing time (TT).

(5.) *Word DVs* – This is the type of simulation that will probably be most useful for anyone interested in running simulations. Selecting this button will generate an output file that contains a variety of the standard dependent measures (e.g., mean first-fixation durations, gaze durations, etc.) for each word in the sentence/text file. With this type of simulation, it is advisable to use a large number of subjects (e.g., at least 100) to obtain stable simulation results. Also, please remember that the predicted results for the first and last words in each sentence do not accurately reflect the model's performance because the model always starts from the middle of the first word (with no parafoveal preview), and because the model always halts (regardless of whatever is happening) when the second stage of lexical processing (i.e., L_2) on the last word has been completed. (For these reasons, the dependent values of the first and last words are never included in our analyses and are coded as zeros in this output; see Reichle, Pollatsek, Fisher & Rayner, 1998). Here is an example of the simulation output and an explanation of what it means. Note that, although the output is displayed across the following six vertically arranged panels (labeled A-F) in this document, in the actual output file the first five types of output are arranged horizontally, with one row per word.

A. The top row lists the dependent variables (means) that are displayed to the right of each word: (1) first-fixation duration (FFD); (2) single-fixation duration (SFD); (3) gaze duration (GD); (4) total time (TT); (5) probability of making one fixation (Pr1); (6) probability of making exactly two fixations (Pr2); (7) probability of making three or more fixations (Pr3+); (8) probability of skipping (PrS); (9) the probability of making an inter-word regression back into a word (PrRI); and (10) the probability of making an inter-word regression out of a word (PrRO).

Sentence	FFD	SFD	GD	TT	Pr(1)	Pr(2)	Pr(3+)	Pr(S)	Pr(RI)	Pr(RO)
--------	---	---	--	--	-----	-----	------	-----	------	-----
Margie	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00
moved	226	228	235	236	0.91	0.04	0.01	0.05	0.01	0.00
into	215	213	228	233	0.75	0.06	0.00	0.19	0.02	0.01
her	213	210	227	230	0.66	0.06	0.00	0.28	0.02	0.02
new	194	189	208	213	0.51	0.05	0.00	0.44	0.01	0.02
apartment	223	232	294	300	0.58	0.21	0.03	0.18	0.03	0.02
at	238	234	240	243	0.62	0.01	0.00	0.38	0.05	0.00
the	200	191	221	230	0.61	0.08	0.00	0.31	0.02	0.05
end	221	220	235	240	0.66	0.06	0.00	0.28	0.02	0.03
of	204	199	215	221	0.44	0.03	0.00	0.53	0.01	0.02
the	187	185	197	198	0.40	0.04	0.00	0.56	0.01	0.01
summer.	0	0	0	0	0.00	0.00	0.00	0.00	0.00	0.00

B. The next set of columns show the initial-fixation landing-site distributions, or the mean proportion of first fixations landing on each of the word's character positions.

Sentence	…	Landing-Site Distributions
--------	…	-------------------------
Margie	…	0.00 0.00 0.00 0.00 0.00 0.00 0.00
moved	…	0.04 0.12 0.21 0.27 0.25 0.11
into	…	0.11 0.15 0.27 0.27 0.20
her	…	0.18 0.21 0.30 0.31
new	…	0.30 0.18 0.25 0.28
apartment	…	0.15 0.07 0.06 0.09 0.12 0.13 0.12 0.09 0.09 0.08
at	…	0.26 0.36 0.38
the	…	0.28 0.25 0.25 0.23
end	…	0.25 0.21 0.26 0.28
of	…	0.43 0.29 0.28
the	…	0.34 0.27 0.22 0.16
summer.	…	0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00

C. The third set of columns show the refixation-probability distributions, or the mean proportion of initial fixations that were followed by a refixation as a function of the location of the initial fixation.

Sentence	…	Refix-Prob Distributions
--------	…	------------------------
Margie	…	0.00 0.00 0.00 0.00 0.00 0.00 0.00
moved	…	0.14 0.13 0.06 0.03 0.04 0.10
into	…	0.43 0.10 0.04 0.01 0.04
her	…	0.37 0.08 0.01 0.02
new	…	0.34 0.11 0.01 0.02
apartment	…	0.71 0.55 0.46 0.35 0.22 0.12 0.14 0.13 0.03 0.02
at	…	0.05 0.05 0.01
the	…	0.39 0.15 0.05 0.05
end	…	0.28 0.10 0.04 0.07
of	…	0.17 0.07 0.03
the	…	0.17 0.08 0.01 0.00
summer.	…	0.00 0.00 0.00 0.00 0.00 0.00 0.00 0.00

D. The fourth set of columns shows the inverted-optimal viewing-position (IOVP) effects, or the durations of single fixations as a function of their location.

Sentence	…	IOVP (1 of 1) Distributions
--------	…	--------------------------
Margie	…	0 0 0 0 0 0 0
moved	…	226 221 231 232 229 218
into	…	211 221 228 220 221
her	…	219 218 220 213
new	…	212 212 200 198
apartment	…	178 230 271 252 244 243 229 248 229 275
at	…	252 240 233
the	…	227 202 200 195
end	…	203 238 235 229
of	…	205 202 200
the	…	190 196 201 196
summer.	…	0 0 0 0 0 0 0 0

E. The fifth set of columns shows the IOVP effects for first-fixation durations, again as a function of their location.

Sentence	…	IOVP (1 of 1+) Distributions									
--------	…	------------------------									
Margie	…	0	0	0	0	0	0	0			
moved	…	210	224	221	236	226	224				
into	…	218	217	224	222	206					
her	…	201	218	218	208						
new	…	194	197	205	191						
apartment	…	185	208	238	239	248	228	220	234	227	245
at	…	253	247	234							
the	…	225	202	202	200						
end	…	215	237	234	219						
of	…	207	189	192							
the	…	185	191	187	181						
summer.	…	0	0	0	0	0	0	0	0		

F. Finally, at the very bottom of the file, several word-based means are shown for each of the five frequency classes of words. These additional measures are formatted as follows:

Frequency-class means:

Class	FFD	SFD	GD	Pr1	Pr2+	PrS
1	244.31	256.19	292.23	0.711	0.182	0.087
2	233.74	241.94	271.73	0.718	0.149	0.118
3	221.49	222.68	246.58	0.716	0.101	0.174
4	212.62	210.06	223.94	0.583	0.039	0.376
5	206.05	202.78	212.71	0.477	0.020	0.502

The seven columns correspond to the following: (1) Class = frequency class (1 = 1-10; 2 = 11-100; 3 = 101-1,000; 4 = 1,001-10,000; 5 = 10,001+); (2) FFD = first-fixation duration; (3) SFD = single-fixation duration; (4) GD = gaze duration; (5) Pr1 = probability of making one fixation; (6) Pr2+ = probability of making two or more fixations; and (7) PrS = probability of skipping.

2. Setting Up Sentence/Text Files

The sentence/text file should contain four columns of information: (1) each word's frequency of occurrence in printed text; (2) each word's (actual) length in character spaces; (3) each word's mean within-sentence predictability (as tabulated by cloze-task experiments); and (4) a copy of the actual word. The last word of each sentence should also be followed by an ampersand (i.e., @), as indicated in the example below. Without this marker, the model program will assume that all of the words in the file are one big sentence, which may or may not be useful. (For more information about the Schilling et al., 1998 sentence corpus, see Reichle et al., 1998.) Here is an example of how the sentence/text file should be formatted:

1	7	0.00	Nancy's
90	7	0.00	kitchen
9816	3	0.40	was
1	8	0.00	infested
7289	4	1.00	with
6	9	0.00	carpenter

1	4	0.90	ants
28850	3	0.80	and
1	8	0.25	roaches.@

The model program should be fairly robust and handle slight variations in formatting (e.g., using blank spaces vs. tabs between columns). However, it's a good idea to make sure that the model is reading in the file correctly using the *WordIVs* mode before you actually run any real simulations. Also, the sentence/text file should be an ascii file (i.e., a file that only contains alphanumeric characters, and no hidden control characters.)

Don't hesitate to contact me if you have any questions or run into any snags. Good luck!

Best regards,
Erik