E-rater: What It Is, How It Works

General Approach

"*E-rater*" is the automated scoring system that has been developed to assist in the evaluation of test takers' responses to open-ended essay questions. The current version of the system is capable of assigning scores (on a 0 to 6 scale) to two kinds of essays -- one that requires examinees to discuss an issue and another that requires them to analyze an argument.

E-rater is empirically based. It is "trained" by being fed samples of essays that have been previously scored by human readers. The samples are randomly selected for each essay prompt so as to represent a complete range of possible scores. Essentially, *e-rater* uses natural language processing techniques to duplicate the performance of human readers.

In its attempt to model human readers, *e-rater* uses several subroutines to extract a wide variety of features of the essays that it evaluates. These features are then used in combination to predict the scores that were assigned previously (by human readers) to the essays in *e-rater's* training sample. Ordinary (stepwise) least squares linear regression analysis is used to select the set of features that best predicts these scores. A total of 50-60 features are "extractable," but in practice only a subset of the most predictive features, usually about 8-12, are retained and used for each essay prompt. The least squares regression weights for these predictors are applied to each new essay to estimate a score. The score is rounded to the nearest integer from 0 to 6 in order to place it on the scale used by human readers.

E-rater is prompt specific. That is, it is trained for each different essay prompt. Thus, the final set of features used to compute scores for any given prompt is likely to differ somewhat across prompts. Several features are much more likely than others to recur as predictors across different prompts. Some of the same features tend to be predictive of scores for both the Issue and the Argument prompt types.

Although *e-rater* is empirically based, it is not "blindly empirical." *E-rater* could, of course, take a "brute-force" approach, extracting numerous features indiscriminately. However, not all of the extractable features correspond equally well to the features that human readers are instructed to consider when they score essays. Therefore, *e-rater* features are required not only to be predictive of human readers' scores, but also to have some logical correspondence to the characteristics that human readers are trained to consider. These characteristics are specified in the scoring rubrics that guide human readers when they score essays.

Specific Techniques

Although certain "surface features" (such as the number of words in an essay) are easily extractable, these purely surface characteristics are *not* used by *e-rater*. Instead, the focus is on three general classes of essay features -- structure (or syntax), discourse (organization), and content (or prompt-specific vocabulary).

Structure

Structural characteristics include such notions as "syntactic variety," that is, the use of various structures in the arrangement of phrases, clauses, and sentences. In this category, several specific features have proven to be predictive of human readers' scores. They are:

- the number of subjunctive modal auxiliary verbs (*would*, *could*, *should*, *might*, *may*)

- the prevalence of infinitive clauses ("*To support his expensive hobby*, Phil Atelist would soon need to get a second job at the post office.")

- the proportion of complement clauses ("He felt *that it was well worth the effort."*)

- the incidence of subordinate clauses ("*Although the argument points out that trends are part of our society*, it does not explain why.").

Organization

How well test takers are able to organize their ideas is a major focus for human readers and for *e-rater*. *E-rater* evaluates organization by extracting a variety of rhetorical features, i.e., characteristics that are associated with the orderly presentation of ideas. Of special interest here are "cue" words or terms that signal where an argument begins and how it is being developed. (By argument, we mean, generally, any rational presentation that uses reasons or examples to persuade the reader.) For example, terms such as "*in summary*" and "*in conclusion*" are used for, what else, summarizing. Words and phrases like the following are used to express opinions or beliefs:

- *certainly, clearly, obviously, plainly, possibly, perhaps, potentially, probably, fortunately, generally, maybe, presumably, unless, albeit, luckily, normally, apparently, herein, likely, surely, ideally, undoubtedly, naturally*

- *for certain, for sure, of course, to some extent, above all, if only, in order to, in order for, so that, so as to*.

The significance of many of the words listed above is that they signal the start of an "argument" (in the sense that we have described), and the number of arguments developed is a feature that is emphasized by *e-rater*. Arguments sometimes begin simply with *I* or *we*, or with "parallel" words or phrases that signify the start of another line of argument. The latter include the following:

> *firstly, essentially, additionally, first, second, third, another, secondly, thirdly, fourth, next, finally, final, last, lastly, moreover, too, also, likewise, similarly, initially, further, furthermore, first of all, in the first place, for one thing, for a start, second of all, many times, more importantly, most importantly*.

The onset of an argument is also signaled by rhetorical words and phrases such as *suppose, supposedly,* and *what if.* The subsequent details are often preceded by words and phrases like:

> *if, specifically, particularly, when, namely, for example, for instance, e.g., in this case, in that case, such that, as well as, in that, such as, about how, in addition, in addition to*.

Words or phrases that are used to contrast points of view or to indicate alternative opinions are often found in arguments. These include words and phrases such as these:

> *otherwise, conversely, however, nonetheless, though, yet, meanwhile, while, but, instead, although, still, notwithstanding, anyway, unlike*, *on the contrary, in contrast, by comparison, in any case, at any rate, in spite of, rather than, on the other hand, even then, even if, even though, apart from, instead of.*

Some words or phrases signal the presentation of evidence to support an argument. These include *since, because, actually, in fact, after all, as a matter of fact,* and *because of.* Still others signify the stage at which inferences are being made:

> *accordingly, consequently, hence, thus, ultimately, so, thereby, then, therefore, following, after, afterward, afterwards, as a consequence, as a result, if so, if not, as such, according to, in turn, right after*.

Yet other words and phrases signal that an argument is being *re*formulated (*alternatively, alternately, that is, in other words, briefly*). In short, E-rater looks for linguistic clues (such as logical connections between sentences and clauses) that signify logical or well-ordered thinking.

Content

By means of "topical analyses," *e-rater* also considers the content or vocabulary of the essays it evaluates. These analyses are predicated on the assumption that well-written essays are more responsive (or relevant) to the topic posed than are poorly written essays. Better essays also tend to use more precise and specialized vocabulary. Therefore, the expectation is that, with respect to the words they contain, good essays will bear a greater resemblance to other good essays than to poorer ones. Weak essays, on the other hand, will be similar to other weak essays.

Acting on this assumption, *e-rater* evaluates each essay and assigns a number based on the similarity of its content to samples of previously-scored essays. This evaluation proceeds as follows. First, for every essay a determination is made of each word's "contribution" to the essay. (Variations of a word, such as "walks," "walking," or "walked," are considered to be the same word.) A word's relative contribution to an essay is estimated by computing a weight for each word in every essay. These word weights reflect both the frequency of a word in a given essay (relative to the frequency of other words in the essay), and the distribution of the word across other essays. The formula is such that words appearing relatively frequently in an essay will, all other things being equal, receive a higher weight than words appearing less frequently in the same essay. However, words that tend to be used in many essays will get lower weights, all else being equal, than those used in few essays.

In order to assign a numerical value to reflect an essay's content, the word weights for each essay are compared with the weights computed for words in the previously scored training essays. An essay's content feature score is assigned by determining the set of essays that its word weights most closely match. Essays whose word weights correlate most strongly with those computed for highly scored essays will get higher content scores than will essays whose word weights resemble those of weaker essays.

As an example, consider the following (partial) results based on essays analyzing an attempt to convince companies to utilize billboards to increase sales. The two most highly weighted words in essays that received the lowest score (i.e., 1) were "picture" and "percent."

Both of these words received much lower weights in essays that got higher scores. On the other hand, the two most heavily weighted words ("recognition" and "local") in the best essays (i.e., 6s) received much lower weights in the poorer essays. The weights assigned to these four specific words in essays at each score level were as follows:

	Score level					
Word	1	2	3	4	5	6
Picture	**.55**	.17	.21	.09	.12	.08
Percent	**.51**	.25	.24	.24	.05	.05
Recognition	.00	.04	.06	.19	.41	**.69**
Local	.05	.11	.11	.20	.16	**.32**

Besides evaluating an essay's content *as a whole*, *e-rater* also considers an essay's content argument-by-argument. The rationale is that additional information can be extracted about an essay's content by examining clusters of word groupings, in this case individual arguments. In a manner similar to that used for essays as a whole, a content score is assigned to each argument on the basis of how well an essay's arguments match the content of essays scored by human readers. The formula used to compute the argument content score assigns higher values to essays with many arguments than to those with few.

E-rater Models

The final step is to use all of the features that *e-rater* extracts (or rather the values that it assigns to these features) to predict the scores assigned by human readers. A model, i.e., a set of features that is most predictive of human readers' scores, is specified for each essay prompt. As stated earlier, the set is generally somewhat different for each prompt. Some features are used much more frequently than others in the predictive models. The most frequently occurring are:

(1) content by argument
(2) content by essay
(3) the number of subjunctive auxiliary verbs
(4) the ratio of subjunctive auxiliary verbs to total words in the essay
(5) the total number of argument development terms.

The weights assigned to these variables may be either positive or negative. For instance, the *number* of subjunctive auxiliary verbs typically receives a positive weight, while the *ratio* of such words is usually weighted negatively. Although the weights assigned to most of the features discussed earlier are usually positive, some features tend to have negative weights. These include certain words used to present evidence (*because, since, actually*, …) and the use of pronouns (*I, we*) to begin arguments.

The overview provided here should give the reader some sense of how *e-rater* functions. More detail is available in a number of reports that can be downloaded from the following Web site: http://www.ets.org./research/erater.html. If you are unable to download, contact us and we'll send the papers electronically as Word documents or by regular mail in hard-copy form.