*ReCALL* **21**(2): 259–279. 2009 © European Association for Computer Assisted Language Learning
doi:10.1017/S0958344009000147 Printed in the United Kingdom

# *Experimenting with a computer essay-scoring program based on ESL student writing scripts*

DAVID CONIAM
*Dept of Curriculum and Instruction, Faculty of Education, The Chinese University of Hong Kong, Sha Tin, Hong Kong*
*(email: coniam@cuhk.edu.hk)*

**Abstract**

This paper describes a study of the computer essay-scoring program BETSY. While the use of computers in rating written scripts has been criticised in some quarters for lacking transparency or lack of fit with how human raters rate written scripts, a number of essay rating programs are available commercially, many of which claim to offer comparable reliability with human raters. Much of the validation of such programs has focused on native-speaking tertiary-level students writing in subject content areas. Instead of content areas with native-speakers, the data for this study is drawn from a representative sample of scripts from an English as a second language (ESL) Year 11 public examination in Hong Kong. The scripts (900 in total) are taken from a writing test consisting of three topics (300 scripts per topic), each representing a different genre. Results in the study show good correlations between human raters' scores and the program BETSY. A rater discrepancy rate, where scripts need to be re-marked because of disagreement between two raters, emerged at levels broadly comparable with those derived from discrepancies between paired human raters. Little difference was apparent in the ratings of test takers on the three genres. The paper concludes that while computer essay-scoring programs may appear to rate inside a 'black box' with concomitant lack of transparency, they do have potential to act as a third rater, time-saving assessment tool. And as technology develops and rating becomes more transparent, so will their acceptability.

Keywords: Assessment, writing, computer scoring, English language

## 1 Introduction

This paper examines a computer essay-scoring program, evaluating how well the program performs on a corpus of already double-marked scripts from an English as a second language (ESL) Year 11 public examination in Hong Kong. The emphasis is on comparisons between the software and human raters so the study does not, therefore, evaluate the program's software *per se*. The scripts are drawn from a writing test consisting of three topics representing three different genres. The program is *BETSY* (the Bayesian Essay Test Scoring sYstem), downloadable for research purposes.[1]

[1] BETSY was funded through the United States Department of Education, and was developed at the College Park of the University of Maryland. It is available as a free download from

The use of computers for assessment purposes has grown considerably since the 1980s, with much research and development in computer-based testing (see Chapelle & Douglas, 2006; Alderson, 2000). Studies demonstrate the advantages of computers in ease of assessment, marking etc. (Alderson, 2000: 594), and in computer adaptive tests using tailored tests (Chalhoub-Deville & Deville, 1999). Nonetheless, the item types produced tend to focus on limited selection types such as multiple-choice (Alderson, 2000: 593). Clapham comments "Until recently, computer testing tended to fossilize existing objective testing methods because objectively marked items such as multiple-choice questions and gap-filling tasks were straightforward to answer on the computer and were easy to mark mechanically" (2000: 7).

In the long term, researchers are aiming to harness the use of natural language processing (NLP) in language assessment. This matches Warschauer and Healey's depiction (1998) of the "Intelligent CALL" phase where software should offer "easy interaction with the material to be learned, including meaningful feedback and guidance" as well as "comprehensible information in multiple media designed to fit the learning style of individual students" (1998: 67). NLP thus represents a long-term aim in assessment, and while elements of NLP are beginning to be incorporated into language tests (see Nerbonne, 2003), it is not yet robust enough for full-scale incorporation into assessment (see Chapelle & Douglas, 2006: 23). More immediately, researchers are investigating how the power of the computer may be harnessed by investigators developing criteria for scoring beyond a simple match. Advancing beyond objectively-scored, essentially right/wrong items, involves computer-assisted response analysis but although there have been advances, this area still remains 'exploratory' (Chapelle & Douglas, 2006: 15).

One area outside the domain of limited-selection item types recently generating interest is the computerised rating of students' written essays (Jamieson, 2005). This area has seen considerable development in recent years, with the following programs now available:

- Intelligent Essay Assessor (Landauer, Foltz & Laham, 1998)
- Project Essay Grade (Page, 2003)
- Bayesian Essay Test Scoring System (BETSY) (Rudner & Liang, 2002)
- E-rater (Educational Testing Service) (Burstein, 2003)

Two major reasons are advanced for using computers to score essays. The first concerns time and money. Access to a reliable computer program may save raters – or indeed teachers – hours grading papers (Chapelle & Douglas, 2006: 34–35). In Hong Kong, The Hong Kong Examinations and Assessment Authority (HKEAA) is mindful of this potential saving with regard to its public examinations. Approximately 80,000 English language writing tests are double marked in a month so tremendous pressure is exerted on raters and the examination system itself. The HKEAA is investigating how certain aspects of the marking process can be computerised (Legislative Council Panel on Education, 2005). Thus, part of the

(http://edres.org/betsy/). The version used in this study was Version 1.03.55d.03.13, produced in February 2003.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

rationale for the current study is to evaluate the program and generate feedback for the HKEAA.

The second claim is that since rating essays is a very subjective task (Hughes, 2003), the use of a computer rater avoids essays being graded or evaluated by human assessors – instead, unbiased and objective software is used. However, validity concerns exist and are discussed below. A further claim by system designers is that the reliability of computerised rating systems matches that of human raters (Dikli, 2006).

While the paper does not focus on how the different computer rating programs function as pieces of software, some background will be provided on computerised essay rating to demonstrate how analysis is conducted.

## 2 Background to computer essay rating

In this section claims for, and the advantages of, computerised essay-rating programs are presented, followed by an examination of criticisms aimed at the software.

### *2.1 Claims for computerised essay-rating programs*

Most essay-grading software functions by analysing sentences and paragraphs, looking for keywords as well as relationships between terms. The Intelligent Essay Assessor (IEA), for example, uses latent semantic analysis (LSA) for the major part of its analysis (Foltz *et al*., 1998). Through training with LSA, a matrix of words and documents is produced. Pieces of writing are then scored to see how well they match the matrix.

The Project Essay Grade (PEG) system, having identified elements such as sentence length, number of paragraphs and elements of punctuation, uses regression to determine how well the different variables correlate with the scores of human raters.

E-rater identifies and examines specific linguistic categories such as grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage. It functions on similar lines to PEG, using regression to predict performance. It compares features in the essays against a bank of sample essays (Attali & Burstein, 2006).

The Bayesian Essay Test Scoring sYstem (BETSY), draws on a broad set of essay features for analysis. After a set of parameters is specified, the system functions by:

- assigning occurrences of content and stylistic features to a set of super-specified levels, according to probability;
- determining the probability of features in input essays falling into the levels stipulated (see Rudner & Liang, 2002).

As BETSY is used in the current study, a more detailed description is presented below.

Several studies have reported favourably on different computerised essay-scoring programs with IEA studies reporting that its scores typically correlated as well with human raters as raters did with each other (Chung & O'Neil, 1997). Foltz *et al.* (1999) report a study with essays written on psycholinguistics by American university students where a correlation of 0.8 was obtained between the program and human raters. Likewise, studies with PEG consistently reported relatively high correlations between PEG and human raters as compared to correlations between human raters

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

(Page *et al*., 1997). Concerning e-rater V.2, Attali and Burstein report correlations with a human rater as high as 0.97 (2006: 22).

In a Graduate Record Examinations (GRE) essays study, Powers *et al*. (2001) concluded that the e-rater system was not ready for use by itself. In high-stakes assessment situations, it needed to be paired with human raters. Powers *et al*. (2002) also report significant, although small, correlations between non-test indicators (self-reported accomplishments and success with various kinds of writing, and self-evaluation of writing ability). Weaker relations emerged with automated scores, however, than with human rater scores.

### *2.2 Criticisms of computerised essay-rating programs*

Against the potential advantages of computerised rating programs and their reliability, counter opinions are now considered.

A major criticism is that the computer rating process is a essentially a "black box" (see Weigle, 2002: 118ff). This same criticism was levelled at "general impression marking" procedures of the 1970s and 80s (Hay, 1982), since rating criteria were not explicit (Weigle, 2002: 112). The assessment of writing by humans has moved, in the past twenty years, from a single impressionistic, holistic scale to scales and descriptors that more clearly define the constructs being assessed. Clearly-defined domain descriptors allow for clearer inferences to be made about test taker abilities. Consequently, computer scoring programs, while returning high correlations with a human rater, still lack transparency as to how they achieve their grades. They are objected to by some teachers of writing. Drechsel (1999), for example, complains that automated essay scoring (AES) systems do not read and understand essays like humans. As Attali and Burstein (2006) comment:

> Whereas human raters may directly evaluate various intrinsic variables of interest, such as diction, fluency, and grammar, in order to produce an essay score, AES systems use approximations or possible correlates of these intrinsic variables (Attali & Burstein, 2006: 3).

Two further issues need to be considered. First, most pieces of writing are currently produced by students writing by hand, i.e., using a pen. If scripts are to be graded by a computer, they need to be machine-readable. This requires clerical assistance to input the scripts. It is not a cost-free option. Secondly, as with speech recognition, a considerable amount of training needs to take place on the program itself before it is operational. In speech recognition systems, users need to train systems to their voice (Coniam, 1998). Likewise, computer rating programs may require hundreds of scripts on a single topic to develop their own potentially reliable rating metric (Dikli, 2006). This initial priming of the program is only feasible with large examination bodies such as ETS, with access to large numbers of scripts. Training is also a time-consuming affair as a representative set of scripts across a range of abilities needs to be available to be fed to the program for training. In the current study, for example, training BETSY with approximately 200 scripts on each of three essay topics (due to the different iterations recommended) took approximately five hours per topic.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Major studies of computerised assessment of written essays have also been skewed: firstly, there has been a focus on native speaker writing; secondly, rating has tended to examine different subject content areas, e.g. using the GRE; thirdly, many have been with tertiary level students. Warschauer and Ware (2006) provide a useful discussion of the place of automated writing evaluation in the ESL classroom; in general, however, there has been little research with secondary ESL students and automated writing assessment. The current study therefore aims to extend the scope of investigation. Firstly, the data is drawn from the writing of secondary school (Year 11, age 17) ESL students. The data comes from the Writing Test of the 2005 Hong Kong Certificate of Education examination (HKCE). Further, as the HKCE Writing Test consists of three topics (each in a different genre), all three are examined.

Data is presented from a number of perspectives. One focus of analysis is the issue of reliability. In addition. the paper compares the program's performance on a range of genres, with some discussion given to a brief qualitative analysis of why certain scripts were over- or under-graded.

To restate, the current study is *not* a validation of a particular computer essay scoring program (i.e., BETSY) *per se*, but rather an investigation into how well such a computerised rating program performs when rating ESL students writing in different genres.

## 3 The current study

This section first provides a short overview of the Hong Kong examination system, the source of the data. It then gives an overview of BETSY. The methodology of text analysis and results retrieval is then outlined. Finally, results are presented and discussed.

### *3.1 The Hong Kong Examination situation*

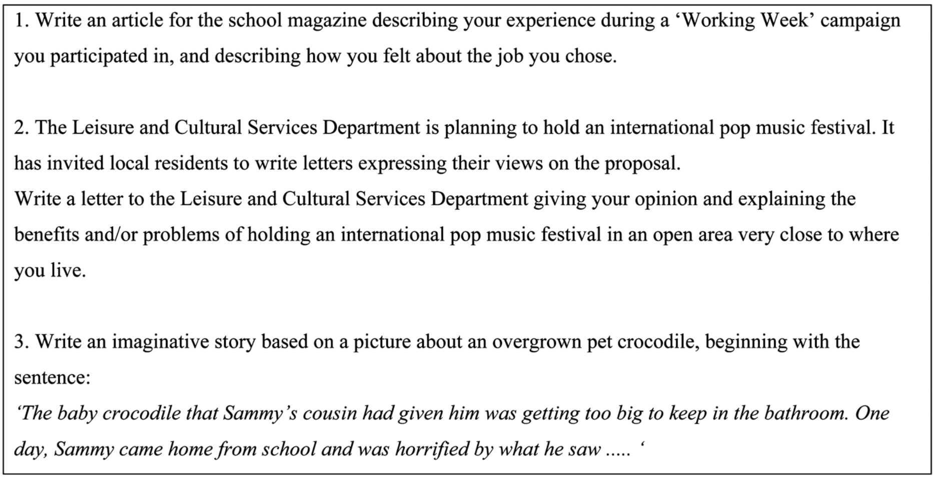
Hong Kong's major public examination is the Hong Kong Certificate of Education (HKCE) examination, administered by the Hong Kong Examinations and Assessment Authority (HKEAA) at the end of Secondary 5 (Grade 11) with a 2005 candidature for English language of 82,078 (HKEAA, 2006). There are four components in the English language HKCE: (1) Writing; (2) Reading; (3) Speaking; and (4) Integrated Reading, Writing and Listening. The 2005 HKCEE Writing paper offered test takers three prompts. Test takers select one prompt and write approximately 300 words within 70 minutes. Overall grades awarded on the HKCE English language paper were A to C (credit), D and E (pass), F and U (Fail).

The HKCE Writing Test has four subscales and descriptors (HKEAA, 2007: 104–105):

1. Relevance and adequacy of content for purpose
2. Accuracy and appropriacy of punctuation, vocabulary, language patterns
3. Planning and organisation
4. Appropriacy of tone, style and register; appropriacy of features for genre.

The subscales each comprise six levels, ranging from 1 (indicating weakness) to 6 (indicating good ability). A level "0" also exists, essentially indicative of "no or minimal performance".

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press



1. Write an article for the school magazine describing your experience during a 'Working Week' campaign you participated in, and describing how you felt about the job you chose.

2. The Leisure and Cultural Services Department is planning to hold an international pop music festival. It has invited local residents to write letters expressing their views on the proposal.
Write a letter to the Leisure and Cultural Services Department giving your opinion and explaining the benefits and/or problems of holding an international pop music festival in an open area very close to where you live.

3. Write an imaginative story based on a picture about an overgrown pet crocodile, beginning with the sentence:
*'The baby crocodile that Sammy's cousin had given him was getting too big to keep in the bathroom. One day, Sammy came home from school and was horrified by what he saw ..... '*

Fig. 1. 2005 HKCE Writing Paper.

Figure 1 (from the 2005 HKCE examination) presents the three prompts: Prompt 1 is descriptive, prompt 2 argumentative and prompt 3 imaginative. Offering a choice from a range of such genres is a longstanding practice in the HKEAA's English language examinations.

## *3.2 BETSY – the Bayesian Essay Test Scoring sYstem*

*3.2.1 Background.* In its automated assessment of student essays, BETSY evaluates essays for style and content. To achieve this, BETSY (as with many automated essay rating systems) first classifies texts according to sets of training materials (Valenti, Neri and Cucchiarelli, 2003). Such classifications include surface features such as: the number of words; average sentence length; number of verbs; content features such as specific words and phrases; and other characteristics including the order in which concepts appear and the occurrence of certain noun-verb pairs.

BETSY operates on Bayesian principles, determining the probability of certain features in an analysed essay being associated with a certain score level. Each essay is considered as a sample of calibrated features such as stemming, stop words, and feature selection. Stemming involves analysing words for key content stems, for example, extracting "educ" from "educate", "education", "educates", "educational", and "educated". Stop words refer to the most frequent articles, pronouns, prepositions in English such as "the", "and", "with". Many of these English high frequency words give little indication of a user's ability; it is an ESL learner's command over the less frequent words that gives an indication of their English language ability (Coniam, 1999). Finally, feature selection involves identification of items with "maximum potential information".

BETSY has three modes of training/analysis: single words (or "key words"); word pairs (or "key phrases"); arguments. "Arguments" are defined by Rudner and Liang (2002: 8) as a pair of words (although not necessarily adjacent) with a prevalence greater than 2% in the essays used for training.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

A score for an essay to be analysed is therefore computed as the product of the probabilities of all the features included in the essay (see Rudner and Liang, 2002, for a more detailed description).

*3.2.2 Operation.* First, users specify how many levels they want BETSY to work with, from two to five. BETSY next needs to be trained with criterial scripts at each defined level. Some programs (IEA, PEG) require only 100 or so scripts for training, but BETSY requires a minimum of two hundred (see Dikli, 2006).

Having selected a mode of training/analysis, the training procedure:

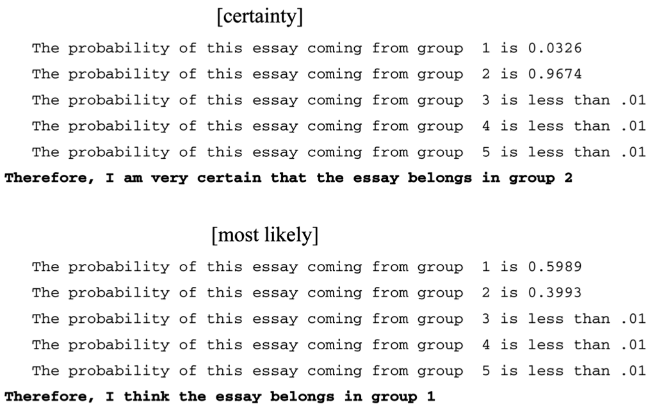
1. trains words
2. evaluates database statistics
3. eliminates uncommon words
4. determines stop words

The process is then reiterated for *word pairs*; and again for *arguments*.

Rudner and Liang (2002: 15) report that "scoring based on arguments tends to outperform scoring based on key words or key phrases". Given Rudner and Liang's stance (although key words and key phrases were also analysed), results are reported for "arguments" ***only*** in the current study because "arguments" subsume key words and key phrases.

*3.2.3 Output.* BETSY produces two major sets of results. The first decides which of three levels a script is assigned: *"certainty"*, *"I think"* and *"I don't know"*. The possibility of a script falling into each of the five criterion levels is presented as a percentage. Where only a single probability is manifestly high, i.e., above 0.9, BETSY declares "certainty". Where there are competing probabilities, BETSY opts for the most likely, declaring that she "thinks". In circumstances where all probabilities are very similar, BETSY declares that she in unable to come to a decision "Therefore I do not know". Figure 2 elaborates.

Secondly, for whatever features have been selected – words, phrases, arguments – BETSY provides tables with the relative probability that a feature would fall into a particular level as against the default possibility of 20% for each level (in the



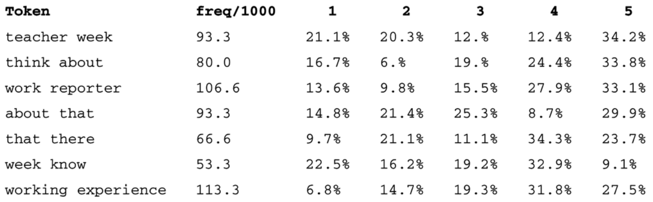
[certainty]

```
   The probability of this essay coming from group  1 is 0.0326
   The probability of this essay coming from group  2 is 0.9674
   The probability of this essay coming from group  3 is less than .01
   The probability of this essay coming from group  4 is less than .01
   The probability of this essay coming from group  5 is less than .01
Therefore, I am very certain that the essay belongs in group 2
```

[most likely]

```
   The probability of this essay coming from group  1 is 0.5989
   The probability of this essay coming from group  2 is 0.3993
   The probability of this essay coming from group  3 is less than .01
   The probability of this essay coming from group  4 is less than .01
   The probability of this essay coming from group  5 is less than .01
Therefore, I think the essay belongs in group 1
```

Fig. 2. BETSY's output – degrees of certainty.



```
Token               freq/1000    1       2       3       4       5
teacher week        93.3         21.1%   20.3%   12.%    12.4%   34.2%
think about         80.0         16.7%   6.%     19.%    24.4%   33.8%
work reporter       106.6        13.6%   9.8%    15.5%   27.9%   33.1%
about that          93.3         14.8%   21.4%   25.3%   8.7%    29.9%
that there          66.6         9.7%    21.1%   11.1%   34.3%   23.7%
week know           53.3         22.5%   16.2%   19.2%   32.9%   9.1%
working experience  113.3        6.8%    14.7%   19.3%   31.8%   27.5%
```

Fig. 3. BETSY's output – relative probability of occurrence of different features.

five-level mode). In Figure 3, level 1 indicates least able, and level 5 most able. Figure 3 presents a sample from Script 240 referred to in Table 10.

### *3.3 Hypotheses*

The working hypotheses in the current study are:

1. Ratings obtained from the program BETSY will not reach interrater correlation levels achieved between "good" human raters. In previous Hong Kong public examinations such as the English language HKCEE, good interrater correlations (i.e., 0.8 or above) are desired.[2]
2. Discrepancies between the computer rater and the human raters will be greater than between human raters.
3. There will be no difference between the levels of correlation reached among the three genres of writing, i.e., genre will not influence scores.

### *3.4 Methodology*

*3.4.1 Preparation/Training*. For each of the three topics, 300 scripts representing a representative cross-section of ability were selected, typed into machine-readable format, and checked for consistency of representation; i.e., that the typing faithfully reproduced test takers' work in spelling, grammar, punctuation and had not been "tidied up".

The scripts were then doubled-marked (using standard HKEAA practice) by nine raters deemed "good" by the HKEAA based on their track record. As good as trained markers may be, however, the use of raw scores to provide accurate information about test takers has been an issue of considerable debate (see McNamara, 1996: 118) as they mask factors such as rater severity and topic effect (Coniam, 2005). Multi-faceted Rasch measurement (MFRM), using the computer program FACETS (Linacre, 1994) helps model and take account of such effects.[3] The double-marked scripts were calibrated onto a Rasch scale to a single score to avoid the problems associated with raw marks

[2] Hatch and Lazaraton (1991: 441) suggest a "strong" correlation, as regards interrater reliability, be taken as 0.8.

[3] In Rasch measurement, the aim is to obtain a unified metric for measurement, with the unit of measurement (referred to as logits) evenly spaced. Logits are centered at zero, zero being the 50% probability represented by an "item" of average difficulty. With a common metric established, different phenomena can be examined and their effects controlled and compared. In principle, this can be achieved independently from situational features associated (as in the current study) with the rating of writing such as prompt difficulty, test taker ability, rater severity levels.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

(Coniam, 2005). Results obtained in MFRM are provided in logits; however, FACETS also provides a "Fair Average" score whereby logit values are converted back onto the original rating scale metric (see Linacre, 1997: 550). The Rasch calibration was therefore taken as the external standard. This calibrated score is henceforth referred to as the "human rater" score; the "Fair Average" score will be referred to as the "level" score, since it makes direct reference to the HKCE six-point scale. By using MFRM, some of the construct-irrelevant variance in test takers' scores, caused by rater strictness and prompt difficulty, is removed. Therefore, the current study moves on from direct comparisons with human raters – as with the Chung and O'Neil (1997), Foltz *et al.* (1999) studies, for example – to comparisons with sets of human scores which may be viewed as more reliable than the raw scores. Thus, the study examines not just how far computer rating correlates with raw scores, but with what may better be considered "true" scores.

The HKEAA uses two measures to examine the quality of raters in its public examinations: interrater correlations; and how many scripts need re-marking because of discrepancies between raters. Analysis of these measures forms a substantial part of the study. In addition to quantitative data, analysis and discussion is presented for some scripts where noticeable differences exist between calibrated human rater scores and BETSY's scores.

The HKEAA also provided, anonymously, test takers' grades for the other components of the HKCE English language examination so comparisons could be made not only between the computerised scoring and the human raters, but between BETSY's scores and the overall HKCE English language examination scores, thus providing an external point of evaluation against BETSY's performance.

Finally textual features such as t-unit length and the number of error-free t-units were computed for each script.[4]

*3.4.2 BETSY setup.* BETSY was specified to operate with five levels of ability, as for HKCE levels: "1" = low and "5" = high ability. Note that analyses are only presented with regard to BETSY's performance in 'argument' mode.

For each topic, the 300 scripts were divided into 200 scripts for training purposes and 100 scripts for subsequent analysis.

## 4 Results and Discussion

This section first presents details of the distribution of scripts and interrater correlations. Raters' performances are then correlated with scripts passed through BETSY. Finally, discussion of individual scripts will explore inappropriate high or low level grades in order to investigate BETSY's decision-making process.

### *4.1 Background analyses – raters and scripts*

The HKEAA's rating scales effectively consist of seven levels, ranging from 0 to 6. Table 1 presents the distribution of test takers' Fair Average scores across levels for the 300 scripts in each topic.

[4] The t-unit as a 'major clausal unit' is generally viewed as providing a more reliable indicator of syntactic complexity than the sentence (see Hunt, 1970).

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Table 1 *Distribution of scripts across Fair Average levels*

| | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|
| Level 0 | 14 | 13 | 9 |
| Level 1 | 49 | 45 | 60 |
| Level 2 | 49 | 51 | 65 |
| Level 3 | 71 | 66 | 67 |
| Level 4 | 68 | 78 | 67 |
| Level 5 | 48 | 47 | 32 |
| Level 6 | 1 | | |

The distribution of scores is generally consistent across the three topics, although Topic 3 is slightly skewed towards the weaker end, with fewer test takers scoring at Level 5. Only one test taker obtained a Level 6 score – on Topic 1.

To provide BETSY with comparative sets of training scripts, the optimal would have been an even spread of Fair Average scores across the five levels. This was not the case, however, as Table 1 indicates; so, to provide BETSY with a representative set of scripts for the five criterion levels, the following procedures were adopted.

Approximately two thirds of the scripts identified at the five levels were selected as training scripts. These scripts were selected from the middle of the level range (where they might be more representative of the level) rather than at the boundary points (where they might be closer to a higher or lower level).

Level 0 scripts were excluded from the training, since Level 0 occurred well below the midpoint of the lowest level, Level 1, as such scripts could skew the definition of Level 1. Level 0 scripts were, however, included for analysis (where they would be expected to be rated as Level 1, the lowest level). The Level 6 script was likewise excluded from the training but included for analysis.

Following the decisions above, the sets of scripts were identified for training and analysis. Table 2 elaborates.

Table 2 shows just under 200 scripts as criterion training scripts for each topic – 188 for Topic 1, 189 for Topic 2 and 193 for Topic 3.

On the HKEAA writing paper, interrater correlations of at least at 0.8 would be expected in terms of good rater reliability [see footnote 2].

Table 3 presents interrater correlations for each of the three prompts. These were produced as the correlation between the total rating (i.e., a maximum of 24 points) given by each rater on the four rating scales, and are presented in Row 1. Data was also made available by the HKEAA for test takers' performance on other components of the examination for comparative purposes (such as their performance on the HKCE examination as a whole). Column 2 presents the correlation between raters and test takers' overall score on the HKCEE examination.

Given the target of 0.8 or better, Table 3 shows acceptable rater levels, indicating a high degree of agreement between them on all three topics. This reliability is confirmed by the raters also correlating highly with the overall subject grade, with correlations in the high 0.8 range or better. All correlations were significant at the 1% level.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Table 2 *Distribution of scripts across levels – training and analysis*

| | | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|---|
| Level 0 | analysis | 14 | 13 | 9 |
| Level 1 | training | 30 | 30 | 40 |
| | analysis | 19 | 15 | 20 |
| Level 2 | training | 33 | 34 | 43 |
| | analysis | 16 | 17 | 22 |
| Level 3 | training | 47 | 45 | 47 |
| | analysis | 24 | 21 | 20 |
| Level 4 | training | 46 | 50 | 46 |
| | analysis | 22 | 28 | 21 |
| Level 5 | training | 31 | 30 | 17 |
| | analysis | 18 | 17 | 15 |
| | Total no. of scripts for training | 188 | 189 | 193 |
| | Total no. of scripts for analysis | 112 | 111 | 107 |
| | | 300 | 300 | 300 |

Table 3 *Interrater correlations*

| | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|
| Rater 1 – Rater 2 correlation | .828 (**) | .883 (**) | .798 (**) |
| Rater – subject mark correlation | .879 (**) | .912 (**) | .858 (**) |

** = $p < .01$.

High interrater correlations are, however, only one measure of reliability. In fairness to test takers, we must consider amounts of disparity between raters for any individual test taker. In assessing writing, re-marking – involving a third rating – is often required when major discrepancies between raters occur. The usual trigger for this in public examination bodies is differences in raters' grades by one subscale level or more.[5]

In the current study, the trigger for re-marking was calculated as more than one point of difference on each of the four scales, i.e., 5 or more points out of 24 (1.25/6). Given this potential level of re-marking between human raters in the current study, Table 4 presents the results for the three topics. The data in the Table indicate the number of scripts that fell within the acceptable level of no more than 4 points of difference out a maximum of 24.

As can be seen, with current raters, re-marking of between 16.7% and 24.3% scripts would have been required, with Topic 1 triggering considerably more re-marking than the other two topics. The results echo those reported by Rudner and

[5] On the GMAT examination in the US, if the two raters differ by more than one score point on the 6-point scale, a third rater remarks the script (Attali and Burstein, 2006: 13).

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Table 4 *Potential amount of remarking required*

| | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|
| Acceptable marking range | 227/300 (75.7%) | 250/300 (83.3%) | 236/300 (78.7%) |
| Remarking required | 24.3% | 16.7% | 21.3% |

Liang (2002: 9) for American high school students writing on a biology topic, where an accuracy rate of 80% was achieved between BETSY and two good raters' results.

We now move to an examination of BETSY's performance with the HKCE Writing Test scripts.

### *4.2 BETSY analyses*

In examining the three possible decisions concerning degree of certainty, Table 5 presents BETSY's decisions on the three writing topics.

BETSY's decisions were either definite or likely, with definite decisions emerging in approximately 80% of the scripts analysed for each topic. No uncertain decisions were recorded.

Table 6 presents the correlations between the "expected" grades (the Fair Average level scores obtained from the two human raters) and the grades awarded by BETSY. Note that in the following tables, results are included for all scripts analysed for the three topics (as in Table 5), irrespective of whether the decision was "I am very certain" or "I think". As all results form part of a rater's statistics in a public examination (that is, no test taker's scores could be omitted because the rater was unsure about the student), the same *modus operandi* is retained.

Table 5 *BETSY's decisions according to degree of certainty*

| Decision | Topic 1 | Topic 2 | Topic 3 |
|---|---|---|---|
| "I am very certain" | 88 (78.6%) | 88 (79.3%) | 88 (80.4%) |
| "I think" | 24 | 23 | 21 |
| "I do not know" | 0 | 0 | 0 |
| N | 112 | 111 | 107 |

Table 6 *Correlations between Fair Average scores and BETSY's scores*

| Correlations | Topic 1 (N = 112) | Topic 2 (N = 111) | Topic 3 (N = 107) |
|---|---|---|---|
| Human rater – BETSY | .826 (**) | .848 (**) | .859 (**) |
| Rater 1 – Rater 2 | .828 (**) | .883 (**) | .798 (**) |

** = p < .01.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Table 6 shows that BETSY achieves strong correlations at the 1% significance level on all three topics. Correlations at the low to mid 0.8 compares very favourably with the interrater correlations achieved between good raters (included from Table 3 above). Topic 2 correlations are slightly lower than that of human raters but Topic 3 has a higher correlation. The results echo those reported by researchers such as Foltz *et al*. (1999) where correlations of 0.8 were obtained between computerised scoring and human raters.

Given that the baseline for re-marking was established earlier (Table 4) as 5 or more points out of 24 (1.25/6), Table 7 below now presents the differences between BETSY's scores and the Fair Average level scores for all scripts that BETSY rated. In Table 7, a positive score indicates that BETSY was more lenient than the level score derived from the two raters; i.e., BETSY rated test takers higher. A negative score indicates that BETSY awarded test takers a lower score. The last row in the Table incorporates interrater discrepancy levels from Table 4.

Table 7, with 1.25 points of difference being taken as criterial for acceptability, shows comparable levels of acceptable marking from BETSY as for two human raters. Topics 1 and 3 are in fact slightly higher than their human discrepancy counterparts, with Topic 2 slightly lower. All three topics are very close to the human results, and further illustrate a lower level of discrepancy. Topic 1 would require most third marking, with Topic 2 again requiring least.

BETSY tends towards stricter rather than more lenient ratings. On Topic 1, the strictest prompt, 5.3% of test takers were rated more than two whole levels lower than by the human rater. Between 1.25 and 2 points of difference, BETSY was between 11.7% and 18.3% stricter. Only a few cases of leniency were recorded – 3.6% of leniency above the 1.25 level on Topic 2. No instances of leniency by more than two levels emerged.

To examine, in greater depth, why BETSY misgraded, each prompt will be discussed in terms of the most overgraded and most undergraded script (see Table 8

Table 7 *Differences between BETSY'S scores and calibrated human rater scores*

| Level of discrepancy | Topic 1 (N = 112) | Topic 2 (N = 111) | Topic 3 (N = 109) |
|---|---|---|---|
| <−2.0 | 6 (5.3%) | 3 (2.7%) | 1 (0.9%) |
| −2.0 – −1.25 | 19 (16.9%) | 13 (11.7%) | 20 (18.3%) |
| −1.24 – +1.24 | 86 **(76.7%)** | 91 **(82.0%)** | 88 **(80.7%)** |
| +1.25 – +2.0 | 1 (0.9%) | 4 (3.6%) | 0 |
| >+2.0 | 0 | 0 | 0 |
| −1.24 to +1.24: human raters | 227/300 **(75.7%)** | 250/300 **(83.3%)** | 236/300 **(78.7%)** |

Table 8 *Most overgraded and undergraded test takers*

| Test taker | Human score | BETSY score | t-units | Error-free t-units |
|---|---|---|---|---|
| 98 | 3.2 | 5 | 35 | 12/35 (34.2%) |
| 240 | 5.2 | 2 | 34 | 32/34 (94.1%) |

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

```
Inner Beauty And Outer Beauty
Being a photographer during the 'Working Week', I have a strong and great feeling on it.
During the 'Working Week', I went to two places. They are differ from each other.
In the first part of the week, I went to a grass field for photoing. The grass field was full of
spirit, surrounded by some tall trees, colourful flower and some active animals. It was such a
divine place in sanscity.
I was captivated by this desirable field. Unconsciously, I forgot my job to take photographs. While
I saw through the len, everything became closer. I can see every movement they taken. So marvellous!
The flying birds was talking to each other!
I lost in comtemplation of admiring the wonderful scenery. I wonder how God can make such a divine
field like that. I am absolutely dote on this piece of grass field!
In the second past of the week, I went to a place which gives me a dreadful feeling in my first
sight all hell broke loose. It was  a low income and standard area. Apast from the people living in
this area, The environment was tumultous. Most of the places were full of rubbish. It was an agony
of me taking photos here. There was not a single step to plesent.
In the course of the week, I stayed with a single family. I comprehended their living condition, it
was uncontrolled for them. They treated me with heart and conscience.
Thought that they live in an unplesent environment, They are kind. I can read through their smile
and eager eyes. We cannot guess what a people alike through Their clothes or other unplesent
factors.
I convinced that I was in wrong atitude towards them at first. I felt very sorry about that.
Through the time being a photographer. My eyes are widen. Every little thing have their own beauty,
just depends on your points of view.
```

```
BETSY's analysis
   The probability of this essay coming from group  1 is less than .01
   The probability of this essay coming from group  2 is less than .01
   The probability of this essay coming from group  3 is 0.0953
   The probability of this essay coming from group  4 is less than .01
   The probability of this essay coming from group  5 is 0.9022

Therefore, I am very certain that the essay belongs in group 5
```

Fig. 4. Overgraded script – Script 98.

below). Since Topic 1 emerged as most divergent on various statistical measures, scripts from this prompt will be examined. Data will be presented for one low ability test taker (scoring below Level 3) who was overrated by BETSY and one high ability test taker (scoring Level 5) who was underrated by BETSY. First, some background textual details of these scripts will be examined, followed by a discussion of BETSY's output.

The most overgraded script was Script 98; it is presented in Figure 4, along with BETSY's certainty rating.

The script above achieved a fair average score of 3.2, with BETSY's rating a "very certain" 5. Against the Level 3 average of 35.3 t-units, this script consisted of 35 t-units, of which 34.2% were free of errors. It did, however, contain many infrequent words, many of which were misspelt (sanscity [presumably, "sanctity"], tumultous, comtemplation, unplesent). While in the training process BETSY supposedly purges infrequently-occurring items, BETSY's analysis is partly predicated on the occurrence of specific words and phrases (see section 3.2). These words may have had an effect on the overall rating. They would not have affected the human raters, who would not have been impressed by misspelt words incorporated into often ungrammatical sentences (reflected in the low percentage of error-free t-units), e.g.,

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Table 9 *Most frequent "arguments" – Script 98*

| token | freq/1000 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| during this | 86.6 | 8.% | 11.6% | 32.% | 9.4% | 39.% |
| that they | 320.0 | 10.8% | 9.8% | 18.5% | 23.7% | 37.2% |
| that being | 260.0 | 3.4% | 9.8% | 17.4% | 33.7% | 35.7% |
| some them | 146.6 | 5.5% | 11.9% | 34.4% | 12.8% | 35.5% |
| they just | 86.6 | 8.% | 17.4% | 13.8% | 28.2% | 32.6% |
| about that | 93.3 | 14.8% | 21.4% | 25.3% | 8.7% | 29.9% |
| from that | 66.6 | 9.9% | 14.3% | 16.9% | 34.8% | 24.1% |
| that there | 66.6 | 9.7% | 21.1% | 11.1% | 34.3% | 23.7% |
| from this | 60.0 | 11.2% | 8.1% | 38.5% | 33.% | 9.1% |
| this time | 73.3 | 26.% | 12.5% | 9.9% | 30.5% | 21.1% |
| that this | 80.0 | 17.1% | 12.4% | 19.6% | 30.1% | 20.8% |
| with them | 260.0 | 6.8% | 22.1% | 19.4% | 29.8% | 22.% |
| take part | 146.6 | 16.8% | 8.1% | 44.9% | 16.5% | 13.7% |
| week have | 146.6 | 16.1% | 23.3% | 33.7% | 9.4% | 17.4% |
| with other | 80.0 | 8.9% | 25.6% | 30.4% | 20.8% | 14.4% |
| make them | 53.3 | 11.2% | 40.6% | 12.8% | 26.3% | 9.1% |
| week felt | 66.6 | 18.7% | 33.8% | 16.% | 16.4% | 15.2% |
| like this | 93.3 | 7.6% | 32.9% | 26.% | 8.9% | 24.6% |
| week which | 60.0 | 38.3% | 20.8% | 21.9% | 11.2% | 7.8% |
| felt about | 60.0 | 37.5% | 20.3% | 16.% | 11.% | 15.2% |
| about this | 93.3 | 34.4% | 24.9% | 11.8% | 12.1% | 16.8% |
| week this | 53.3 | 30.7% | 14.8% | 17.5% | 12.% | 24.9% |
| during which | 86.6 | 29.9% | 21.7% | 21.4% | 8.8% | 18.2% |
| that time | 213.3 | 29.1% | 15.8% | 16.6% | 14.9% | 23.6% |
| this first | 66.6 | 18.8% | 20.4% | 16.1% | 22.% | 22.8% |
| what they | 180.0 | 18.3% | 6.6% | 23.5% | 29.4% | 22.2% |
| have some | 53.3 | 11.3% | 24.5% | 19.4% | 26.5% | 18.3% |

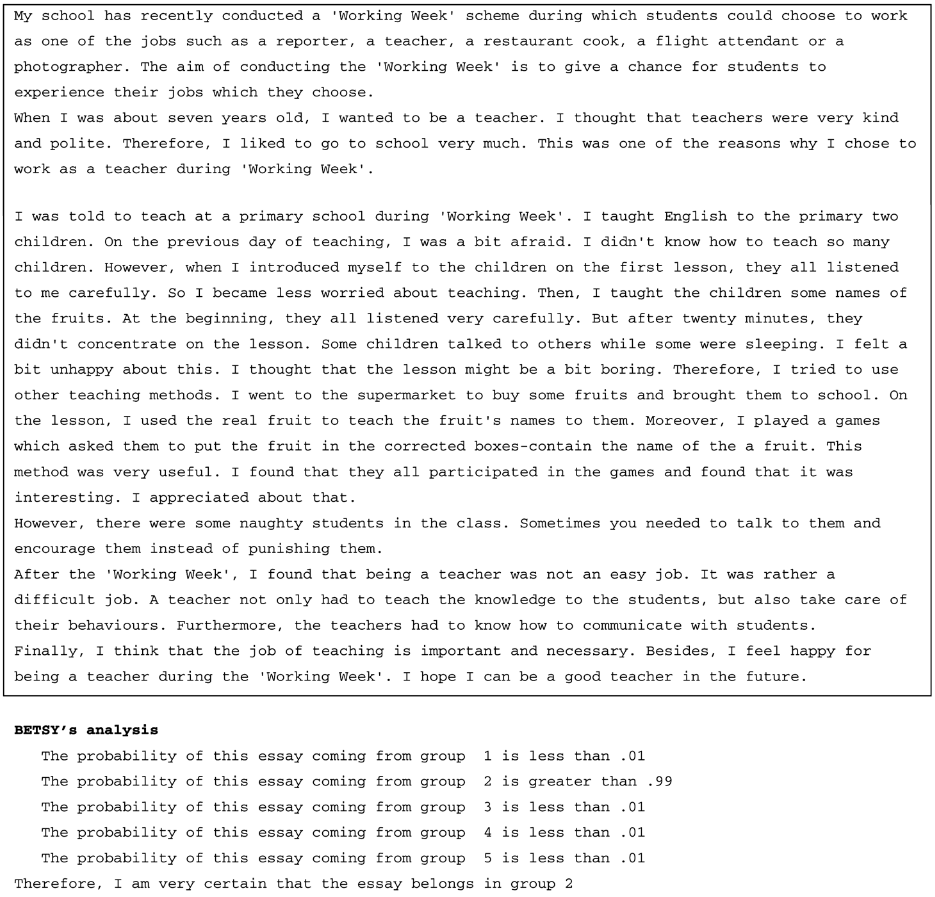
"Thought that they live in an unplesent environment, They are kind". It is a possible indicator, nonetheless, as to how computerised rating programs may err, and although it is not the focus of the current study, one issue which could be investigated further in computerised rating is the extent to which a computer may be fooled by writers (Page, 2003).

Table 9 presents a sample of the output produced by BETSY for Script 98. For each of the five levels ("5" indicating most able), BETSY provides the possibility that a particular "argument" is indicative of a particular level. In Table 9, frequencies above 30% (as against the average of 20%) have been bolded to highlight where the highest amount of decision-pointing frequencies occur among the five levels.

As Table 9 illustrates, a higher percentage of frequencies above 30% occur in the Level 4 and 5 columns even though the data in the Table essentially contains collocations between function words, and as such is not particularly informative. It is presumably partly for this reason that BETSY is "very certain" that this is a Level 5 script.

In contrast with an overgraded script, the most undergraded script is now presented – Script 240 in Figure 5. Against the Level 5 average of 41.7 t-units, Script 240

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press



My school has recently conducted a 'Working Week' scheme during which students could choose to work as one of the jobs such as a reporter, a teacher, a restaurant cook, a flight attendant or a photographer. The aim of conducting the 'Working Week' is to give a chance for students to experience their jobs which they choose.

When I was about seven years old, I wanted to be a teacher. I thought that teachers were very kind and polite. Therefore, I liked to go to school very much. This was one of the reasons why I chose to work as a teacher during 'Working Week'.

I was told to teach at a primary school during 'Working Week'. I taught English to the primary two children. On the previous day of teaching, I was a bit afraid. I didn't know how to teach so many children. However, when I introduced myself to the children on the first lesson, they all listened to me carefully. So I became less worried about teaching. Then, I taught the children some names of the fruits. At the beginning, they all listened very carefully. But after twenty minutes, they didn't concentrate on the lesson. Some children talked to others while some were sleeping. I felt a bit unhappy about this. I thought that the lesson might be a bit boring. Therefore, I tried to use other teaching methods. I went to the supermarket to buy some fruits and brought them to school. On the lesson, I used the real fruit to teach the fruit's names to them. Moreover, I played a games which asked them to put the fruit in the corrected boxes-contain the name of the a fruit. This method was very useful. I found that they all participated in the games and found that it was interesting. I appreciated about that.

However, there were some naughty students in the class. Sometimes you needed to talk to them and encourage them instead of punishing them.

After the 'Working Week', I found that being a teacher was not an easy job. It was rather a difficult job. A teacher not only had to teach the knowledge to the students, but also take care of their behaviours. Furthermore, the teachers had to know how to communicate with students.

Finally, I think that the job of teaching is important and necessary. Besides, I feel happy for being a teacher during the 'Working Week'. I hope I can be a good teacher in the future.

**BETSY's analysis**

```
   The probability of this essay coming from group  1 is less than .01
   The probability of this essay coming from group  2 is greater than .99
   The probability of this essay coming from group  3 is less than .01
   The probability of this essay coming from group  4 is less than .01
   The probability of this essay coming from group  5 is less than .01
Therefore, I am very certain that the essay belongs in group 2
```

Fig. 5. Undergraded script – Script 240.

consisted of 34 t-units, fewer even than Script 98 with its 35 t-units. However, Script 240 contained 94.1% error-free t-units.

In contrast to Script 98, where the test taker attempted to produce elaborate structures with flowery vocabulary beyond their ability level (as evidenced by the low number of error-free t-units), the writer of Script 240 writes in a style containing simple, mainly accurate structures, using vocabulary that shows a certain, although limited, range. One feature that BETSY examines is sentence length, which is comparatively shorter in Script 240, a possible reason for the Level 2 score.

There is little low-frequency vocabulary in Script 240. The style is simpler and less flowery than script 98. Consequently, whereas Script 240 achieved a human level score of 5.2, BETSY was "very certain" that the script was a "2". Table 10 presents the most frequent arguments.

Script 98's arguments consisted only of function words whereas the arguments in Script 240 show many more content word collocations, although substantially more occur in levels 1 and 2 than in the upper levels. It is, presumably, on this basis that the script has emerged as a "2".

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Table 10 *Most frequent "arguments" – Script 240*

| Token | freq/1000 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| teacher week | 93.3 | 21.1% | 20.3% | 12.% | 12.4% | 34.2% |
| think about | 80.0 | 16.7% | 6.% | 19.% | 24.4% | 33.8% |
| work reporter | 106.6 | 13.6% | 9.8% | 15.5% | 27.9% | 33.1% |
| about that | 93.3 | 14.8% | 21.4% | 25.3% | 8.7% | 29.9% |
| that there | 66.6 | 9.7% | 21.1% | 11.1% | 34.3% | 23.7% |
| week know | 53.3 | 22.5% | 16.2% | 19.2% | 32.9% | 9.1% |
| working experience | 113.3 | 6.8% | 14.7% | 19.3% | 31.8% | 27.5% |
| that this | 80.0 | 17.1% | 12.4% | 19.6% | 30.1% | 20.8% |
| with them | 260.0 | 6.8% | 22.1% | 19.4% | 29.8% | 22.% |
| students also | 80.0 | 17.1% | 24.8% | 34.2% | 10.% | 13.9% |
| told them | 113.3 | 13.2% | 14.3% | 34.% | 11.6% | 26.8% |
| experience that | 53.3 | 11.2% | 16.2% | 32.1% | 13.2% | 27.3% |
| with other | 80.0 | 8.9% | 25.6% | 30.4% | 20.8% | 14.4% |
| teacher when | 93.3 | 22.8% | 16.5% | 30.4% | 17.9% | 12.4% |
| very important | 60.0 | 20.2% | 36.5% | 17.3% | 17.8% | 8.2% |
| during lesson | 100.0 | 13.9% | 35.3% | 19.9% | 8.2% | 22.6% |
| week felt | 66.6 | 18.7% | 33.8% | 16.% | 16.4% | 15.2% |
| teacher after | 53.3 | 11.1% | 32.2% | 19.1% | 19.6% | 18.1% |
| after lesson | 53.3 | 10.8% | 31.4% | 12.4% | 19.1% | 26.4% |
| week chose | 53.3 | 30.6% | 29.5% | 17.5% | 6.% | 16.5% |
| felt about | 60.0 | 37.5% | 20.3% | 16.% | 11.% | 15.2% |
| choose work | 380.0 | 35.1% | 12.7% | 18.8% | 19.3% | 14.2% |
| about this | 93.3 | 34.4% | 24.9% | 11.8% | 12.1% | 16.8% |
| teacher restaurant | 233.3 | 33.3% | 24.1% | 9.5% | 19.5% | 13.5% |
| that school | 100.0 | 32.7% | 14.2% | 11.2% | 15.3% | 26.5% |
| week this | 53.3 | 30.7% | 14.8% | 17.5% | 12.% | 24.9% |
| cook flight | 206.6 | 30.2% | 24.6% | 13.% | 19.9% | 12.3% |
| week school | 86.6 | 30.1% | 27.2% | 17.2% | 13.2% | 12.2% |
| during which | 86.6 | 29.9% | 21.7% | 21.4% | 8.8% | 18.2% |
| conducted working | 406.6 | 29.5% | 22.9% | 18.1% | 21.% | 8.6% |
| teacher working | 60.0 | 29.3% | 21.2% | 22.3% | 11.4% | 15.8% |
| very easy | 60.0 | 29.2% | 21.1% | 11.1% | 22.8% | 15.8% |
| could teach | 66.6 | 28.3% | 20.5% | 27.% | 16.6% | 7.7% |
| very difficult | 86.6 | 23.6% | 11.4% | 18.% | 27.7% | 19.2% |
| think that | 340.0 | 22.8% | 22.% | 23.1% | 17.8% | 14.4% |
| experience during | 193.3 | 21.% | 27.4% | 24.% | 17.3% | 10.2% |
| students were | 280.0 | 20.7% | 12.9% | 23.7% | 13.9% | 28.8% |
| first lesson | 106.6 | 20.4% | 19.7% | 19.4% | 23.9% | 16.6% |
| teach class | 60.0 | 20.3% | 22.% | 17.4% | 23.8% | 16.5% |
| scheme school | 60.0 | 20.1% | 14.6% | 17.2% | 23.6% | 24.5% |
| good teacher | 166.6 | 19.% | 24.1% | 19.% | 22.3% | 15.5% |
| know that | 166.6 | 19.% | 20.6% | 19.% | 22.2% | 19.2% |
| this first | 66.6 | 18.8% | 20.4% | 16.1% | 22.% | 22.8% |
| teacher during | 80.0 | 17.3% | 18.7% | 24.7% | 25.3% | 14.% |
| when they | 133.3 | 17.1% | 12.4% | 22.8% | 20.% | 27.7% |
| week scheme | 1126.6 | 11.6% | 16.8% | 19.5% | 24.4% | 27.7% |
| teacher easy | 226.6 | 11.5% | 25.% | 26.3% | 24.8% | 12.5% |
| teacher future | 60.0 | 10.5% | 15.2% | 24.% | 24.7% | 25.6% |
| them after | 73.3 | 9.3% | 27.% | 26.6% | 21.9% | 15.1% |

## 5 Conclusion

This paper has investigated the use of a computer essay scoring program, examining the performance of Year 11 secondary school ESL students on a Hong Kong public examination. The data included the three different prompts from the HKCE Writing Test to see whether there might be any effect due to the prompt's genre.

The study was based on three hypotheses. The first was that ratings obtained from the computer rating program BETSY would not reach interrater correlation levels achieved between "good" human raters. This hypothesis was therefore rejected as correlations were generally comparable with the human rater score. Results showed good correlations between a calibrated score based on two good human raters' scores and the program BETSY, and are comparable with other researchers' findings (e.g., Rudner and Liang, 2002).

The second hypothesis was that there would be greater discrepancies between the computer rater and the human raters than between human raters. This hypothesis was also rejected since overall levels of discrepancy were very similar to those obtained by the human raters. The rater discrepancy rate, where scripts need to be re-marked because of disagreement between two raters, emerged at levels broadly comparable with those derived from discrepancies between paired human raters. One issue worthy of further research is why more discrepancies were due to BETSY rating more harshly rather than leniently.

The third hypothesis was that there would be no difference between the levels of correlation reached among the three genres. This hypothesis was also rejected since, while the three topics in the study were generally comparable, the descriptive topic emerged on a number of issues as being the most problematic. Among these, correlations with the human rater was the lowest of the three genres. Also it produced the largest number of discrepancy scripts requiring most re-marking. Further, the descriptive topic emerged as the strictest in terms of lowest grades awarded to test takers, and recorded (by a small measure) the least number of "most certain" decisions. This might be an area of future research – possibly investigating these issues further in terms of linguistic features such as range of vocabulary, structures and discourse patterns associated with this genre.

In summary then, while computer rating programs have their detractors in terms of transparency, it can be seen that they produce results which compare favourably with human raters. Where resources are limited and the examination is low stakes, a computer rating system could be used as a second marker, with a third rater invoked when large discrepancies are flagged.

Despite high correlations with a human rater, computer programs still lack transparency in terms of how they achieve the grades they award. This creates a validity problem as we do not know how ratings are achieved, e.g. BETSY's definition of an "argument" when two content words collocate at 2% of the overall text, irrespective of punctuation or other features. This definition appears illogical, unclear, and unargument-like, representing a problem for teachers of writing (Weigle, 2002: 236) in that it too much resembles a "black box". We must nonetheless note that with the "general impression marking" schemes of the 1980s criteria were not always explicit either. Human rating schemes have not always been transparent.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Currently, in Hong Kong public examinations, essays are not typed. As a computer rating system entails creating a database for training and analysis, scripts would need to be scanned in. This is, at present, a major exercise. However, college and high school students sometimes submit typed written assignments. In this case, if large numbers of students produce scripts at the same time at the end of the school year, for example, a computer-based system might soon be a viable assessment tool. It could also have uses when large number of applicants are applying to a college, for example, and their writing needs to be quickly analysed.

Since many religious organisations or trade unions operate a number of schools in Hong Kong, their schools could cooperate by setting similar writing topics for their end-of-year examinations. Once a large data bank is created and the system trained adequately, teachers could single mark scripts (the norm is double marking), using the computer as the second rater, provided arrangements for a third rater exist for discrepancies.

In this study, computer rating systems have been examined from an assessment perspective but other uses also exist. Warschauer and Ware (2006), for instance, discuss the potential of automated writing evaluation in the classroom, and Cheung *et al.* (2005) discuss a system based on latent semantic analysis for formative feedback when student essays are compared against models produced by the teacher. So while it may still be early days for computerised essay-scoring programs – at least with the rating of second-language writing – the potential for computerised essay-scoring cannot be ignored, as improvements develop in this area over the coming decade.

The study's limitations must also be considered. The first issue is sample size, with 200 training scripts being close to the threshold of tolerance for reliability. Also, while the study has investigated what computer essay-rating programs are capable of, it has not attempted to validate BETSY *per se*. Indeed, advances in computing and NLP have begun to surpass the technology underpinning the possibly ageing BETSY (the version used in this study dates from 2003). Researchers such as Attali and Burstein (2006: 25) discuss the potential of automated essay rating systems to develop ''an objective writing scale that is independent of specific human rubrics and ratings''. Interesting developments await on the computer rating horizon.

## Acknowledgement

I would like to thank Prof. Lawrence Rudner for making BETSY accessible as a free download, and the Hong Kong Examinations and Assessment Authority for the HKCE English language scripts, thereby making it possible to experiment with BETSY.

## References

Alderson, J. C. (2000) Technology in testing: The present and the future. *System*, **28**: 593–603.

Attali, Y. and Burstein, J. (2006) Automated essay scoring with e-rater v.2. *Journal of Technology, Learning, and Assessment*, (*JTLA*) **4**(3).

Burstein, J. (2003) The e-rater scoring engine: Automated essay scoring with natural language processing. In: Shermis, M. D. and Burstein, J. (eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 113–122.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Chalhoub-Deville, M. and Deville, C. (1999) Computer adaptive testing in second language contexts. *Annual Review of Applied Linguistics*, **19**: 273–299.

Chapelle, C. A. and Douglas, D. (2006) *Assessing Language through Computer Technology*. Cambridge, UK: Cambridge University Press.

Cheung, W. K., Mørch, A. I., Wong, K. C., Lee, C., Liu, J. and Lam, M. H. (2005) Grounding Collaborative Learning in Semantics-Based Critiquing. In: Rynson, W. H., Lau Qing Li, Cheung, R. and Wenyin Liu (eds.), *Advances in Web-based learning – ICWL 2005*. New York: Springer, 244–255.

Chung, G. K., and O'Neil, H. F., Jr. (1997) Methodological Approaches to Online Scoring of Essays. ERIC Document Reproduction Service No. ED 418 101.

Clapham, C. (2000) Assessment and testing. *Annual Review of Applied Linguistics*, **20**: 147–161.

Coniam, D. (2005) Raw scores as examination results: How far can they be relied upon? Paper presented at the ALTE Second International Conference, Berlin, 19–21 May 2005.

Coniam, D. (1998) Voice recognition software accuracy with second language speakers of English. *System*, **27**(1): 1–16.

Coniam, D. (1999) Second language proficiency and word frequency in English. *Asian Journal of English Language Teaching*, **9**: 59–74.

Dikli, S. (2006) An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, **5**(1). http://escholarship.bc.edu/jtla/

Drechsel, J. (1999) Writing into Silence: Losing Voice with Writing Assessment Technology. *Teaching English in the Two-Year College*, **26**(4): 380–387.

Foltz, P. W., Kintsch, W. and Landauer, T. K. (1998) The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes*, **25**(2&3): 285–307.

Foltz, P. W., Laham, D. and Landauer, T. K. (1999) The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, **1**(2). http://imej.wfu.edu/articles/1999/2/04/index.asp

Hay, J. (1982) General Impression Marking: Some Caveats. *English in Australia*, **59**: 50–58.

Hatch, E. and Lazaraton, A. (1991) *The Research Manual: Design and Statistics for Applied Linguistics*. Boston, MA: Heinle and Heinle.

Hong Kong Examinations and Assessment Authority (HKEAA) (2006) *Language Proficiency Assessment for Teachers (English Language) 2006: Assessment Report*. http://eant01.hkeaa.edu.hk/hkea/redirector.asp?p_direction=body&p_clickurl=otherexam%5Fbycategory%2Easp

Hong Kong Examinations and Assessment Authority (HKEAA) (2007) *HKCEE English language examination report and question papers*. Hong Kong: Hong Kong Examinations and Assessment Authority.

Hughes, A. (2003) *Testing for language teachers*. Cambridge, UK: Cambridge University Press.

Hunt, K. W. (1970) Syntactic maturity in school children and adults. *Monographs of the Society for Research in Child Development*, 135(35/1). Chicago: University of Chicago Press.

Jamieson, J. (2005) Research in language assessment: trends in computer-based second language assessment. *Annual Review of Applied Linguistics*, **25**: 228–242.

Landauer, T. K., Foltz, P. W. and Laham, D. (1998) Introduction to Latent Semantic Analysis. *Discourse Processes*, **25**: 259–284.

Legislative Council Panel on Education (2005) LC Paper No. CB(2)323/05-06(01): Grant to support the modernization and development of the Hong Kong Examinations and Assessment Authority's examination systems. http://www.legco.gov.hk/yr05-06/english/panels/ed/papers/ed1114cb2-323-1e.pdf

Linacre, J. M. (1994) *FACETS: Rasch Measurement Computer Program*. Chicago: MESA Press.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press

Linacre, J. M. (1997) Communicating examinee measures as expected ratings. *Rasch Measurement Transactions*, **11**(1): 550–551. Retrieved October 11, 2007, from http://www.rasch.org/rmt/rmt111m.htm

Nerbonne, J. (2003) Natural Language Processing in Computer-Assisted Language Learning. In: Mitkov, R. (ed.), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 670–698.

McNamara, T. (1996) *Measuring second language performance*. New York: Longman.

Page, E. B. (2003) Project Essay Grade: PEG. In: Shermis, M. D. and Burstein, J. (eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, 43–54.

Page, E. B., Poggio, J. P. and Keith, T. Z. (1997) Computer analysis of student essays: Finding trait differences in the student profile. AERA/NCME Symposium on Grading Essays by Computer.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E. and Kukich, K. (2001) *Stumping e-rater: Challenging the validity of automated essay scoring* (GRE Board Professional Rep. No. 98–08bP, ETS Research Rep. No. 01–03). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M. S., Fowles, M. E. and Kukich, K. (2002) Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research*, **26**(4): 407–425.

Rudner, L. M. and Liang, T. (2002) Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment*, **1**(2). http://www.jtla.org

Valenti, S., Neri, F. and Cucchiarelli, A. (2003) An overview of current research on automated essay grading. *Journal of Information Technology Education*, (2), 319–330.

Warschauer, M. and Healey, D. (1998) Computers and language learning: An overview. *Language Teaching*, **31**: 57–71.

Warschauer, M. and Ware, P. (2006) Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, **10**(2): 1–24.

Weigle, S. C. (2002) *Assessing writing*. Cambridge: Cambridge University Press.

https://doi.org/10.1017/S0958344009000147 Published online by Cambridge University Press