UNIVERSITY OF MURCIA

International Journal of English Studies

IJES

www.um.es/ijes

The Impact of Computer-Based Feedback on Students' Written Work

KHALED EL EBYARY

Newcastle University, UK and Alexandria University, Egypt

SCOTT WINDEATT

Newcastle University, UK

Received: 30 July 2010 / Accepted: 22 October 2010

ABSTRACT

While research in second language writing suggests that instructor feedback can have a positive influence on students' written work, the provision of such feedback on a regular basis can be problematic, especially with larger student numbers. A number of computer programs that claim to provide both automatic computer-based holistic scores and computer-based feedback (CBF) on written work are available and therefore have the potential to deal with this issue. *Criterion* is one such tool that claims to be able to provide automated feedback at word, sentence, paragraph and text level, but there is still a need for more research into the practical value of providing feedback on L2 writing. Quantitative and qualitative data about feedback practice was collected from 31 instructors and 549 Egyptian trainee EFL teachers using pre-treatment questionnaires, interviews and focus groups. 24 of the trainees then received computer-based feedback using *Criterion* on two drafts of essays submitted on each of 4 topics. Data recorded by the software suggested a positive effect on the quality of students' second drafts and subsequent submissions, and post-treatment questionnaires, interviews and focus groups showed a positive effect on the students' attitudes towards feedback.

KEYWORDS: computer-based feedback, *Criterion*, attitudes and motivation

RESUMEN

A pesar de que la investigación sobre escritura en segundas lenguas sugiere que los comentarios de los profesores pueden tener una influencia positiva sobre el trabajo escrito de los estudiantes, el proporcionar con regularidad tales comentarios puede ser problemático, especialmente en clases muy numerosas. Sin embargo, existen en el mercado una serie de programas informáticos que garantizan poder proporcionar tanto evaluaciones integrales de carácter automático como comentarios informatizados sobre trabajos escritos y que, por lo tanto, tienen cierto potencial para tratar este problema. *Criterion* es una de estas herramientas y, como tal, proporciona información automatizada a nivel de palabra, oración, párrafo y texto. En el presente trabajo analizamos el valor práctico que ofrece en la producción de comentarios para la escritura en L2 y, a este respecto, recogimos datos cuantitativos y cualitativos de 31 instructores y 616 profesores en formación de inglés como lengua extranjera de origen egipcio por medio de cuestionarios previos, entrevistas y discusiones en grupo. 24 de los profesores recibieron comentarios informatizados producidos por medio de *Criterion* sobre dos borradores de redacciones realizadas acerca de 4 temas diferentes. La información registrada en el software indica un efecto positivo sobre la calidad de los segundos borradores realizados por los estudiantes, así como de escritos posteriores. Asimismo, tanto los cuestionarios administrados después de la aplicación, como las entrevistas y las discusiones en grupo revelan un efecto positivo sobre la actitud de los estudiantes hacia los comentarios de los profesor

PALABRAS CLAVE: feedback por ordenador, *Criterion*, actitudes, motivación

Address for correspondence: Khaled El Ebyary, Curricula and Methods of Teaching Department, School of Education, Damanhour, Alexandria, Egypt. Tel.: +2-0192666817 email: k.ebyary@newcastle.ac.uk

ISSN: 1578-7044

1. INTRODUCTION

One important assumption in feedback research is that the provision of feedback can improve learning (Black and Wiliam, 1998a; Black and Wiliam, 1998b; Gibbs and Simpson, 2004; Hyland and Hyland, 2006). Research results however, are inconclusive with respect to the most useful focus of feedback comments on L1 and L2 learners' writing (e.g. grammar, lexis or organisation/structure), the form in which they are given (e.g. explicit or indicative), and the source of the feedback (i.e. instructors or peers). Furthermore, there is evidence that the quality of feedback students receive on assessed work remains a widespread source of dissatisfaction, and that the level of satisfaction may even be declining (Huxham, 2007). Feedback on writing is a time-consuming task for instructors because they may not be able to give individualized, immediate, content-related feedback to multiple drafts (Grimes and Warschauer, 2010; Lee et al, 2009). In addition, with most feedback practices students tend to be considered as mere recipients, leading some (e.g. Ferris, 2003; Lee, 2007) to describe existing practice as more teacher-centred in the sense that the focus is on teachers' actions rather than on students' reactions. Indeed, although researchers emphasize that feedback is meant to benefit students as it offers the type of '*individualized attention that is otherwise rarely possible under normal classroom conditions*' (Hyland and Hyland, 2006,p.xv), ensuring that feedback is provided on a regular basis can be problematic, especially with large student numbers. Peer feedback can have a role to play, but research suggests that learners see this as having a different purpose from instructor feedback (Jacobs et al., 1998), and finding additional teachers to provide feedback is often impractical. One possible solution is to use computer applications that can generate feedback, or "intelligent CALL". In Warschauer and Healey's (1998) words, 'intelligent CALL' refers to computer applications which can interact with '*the material to be learned, including (providing) meaningful feedback and guidance*'. Warschauer and Ware (2006) summarise some of the features of three such applications - (see table 1).

Company	Software Engine	Evaluation Mechanism	Commercial Product	Scoring	Feedback
Vantage learning	Intellimetric	Artificial intelligence	MY Access	Holistic and component scoring	Limited individualized feedback
ETS	E-rater and Critique	Natural language processing	Criterion	Single holistic score	Wide range of individualized feedback
Pearson Knowledge Technologies	Intelligent Essay assessor	Latent semantic analysis	Holt Online Essay Scoring (and others)	Holistic and component scoring	Limited individualized feedback

Table 1. Comparison of major AES systems (Warschauer & Ware, 2006)

Research on such applications has generally focused on their use for assessing writing rather than on providing feedback (e.g. Rudner and Liang, 2002), comparing human scoring to computer scoring (e.g. Wang and Brown, 2007), and validating computerised scoring systems (e.g. Powers et al., 2001), and claims have been made that the reliability of such applications in assessing writing matches that of human raters (e.g. Dikli, 2006). While some investigations have been carried out into the usefulness of such applications in generating computer-based feedback (CBF) on students' written work (e.g Attali 2004; Coniam, 2009) there are still relatively few research studies in this area. To address this gap in the research, this paper reports a study which investigates the effect of providing computer-based feedback using *Criterion* on the attitudes of a particular group of students towards feedback, and on their writing process and product.

2. LITERATURE REVIEW

2.1. Computer-based feedback

There are a number of computer applications which evaluate and score written work, some of which also provide formative feedback to the writer. Such applications are known as *Automated Essay Scoring* (AES) (Shermis & Buretein, 2003) or *Automated Writing Evaluation* (AWE) (Warschauer & Ware, 2006) systems, and examples include *e-rater*, *MY Access*, *Holt Online Scoring*, *BETSY* and *Criterion*. AES (or AWE) has been described as computer technology that evaluates and scores written prose with the purpose of saving time, reducing cost, and increasing reliability, in the assessment of writing (Chung and O'Neil, 1997; Hamp-Lyons, 2001; Page, 2003; Rudner and Liang, 2002; Shermis and Barrera, 2003; Shermis and Burstein, 2003; Rudner and Cagne, 2001; Warschauer and Ware, 2006). Such software has been developed to score students' writing in a variety of genres, to generate numerical scores and in some cases to provide other forms of feedback (Warschauer and Ware, 2006).

Coniam (2009) summarizes the major arguments in the literature (e.g. Chapelle & Douglas, 2006; Dilki, 2006; Hughes, 2003) for using computers in assessing students' written work as money, time, objectivity, and reliability levels matching those attained by multiple human raters. Bull and McKenna (2004) argue that the use of computers in assessing written responses is pedagogically desirable as it can be integrated with existing assessment methods and strategies, increase the frequency of feedback, and broaden the range of assessed skills.

2.2. The effectiveness of AES/AWE applications

However, the research into the use of AES/AWE applications paints a confusing picture, with some studies reporting favourable results (Coniam, 2009; Hutchison, 2007) while others

report negative or mixed results (Lai 2010; Lee et al, 2009; Tuzi, 2004), with factors such as individual writing ability, the pedagogy adopted and the particular AES/AWE application influencing the results (Lee et al, 2009). For example, less trained writers faced difficulties in using revision tools (Kozna and Johnston, 1991); learners using My Access were dissatisfied with the grade the software awarded them, and with both the accuracy and clarity of feedback on content and the rhetorical aspects of their writing (Chen and Cheng, 2006). In contrast, a number of case studies (e.g. Dmytrenko-Ahrabian, 2008; Ellison, 2007; Ussery, 2007) provide reports, though only anecdotal, of student and teacher satisfaction with the *Criterion* software.

In terms of AES effectiveness in scoring written work as opposed to providing formative feedback, Coniam (2009) claims that BETSY awarded scores for exam scripts written by Year 11 ESL students in Hong Kong that were broadly comparable with those awarded by human raters. Hutchison (2007) reported similarly positive results in a study of essays written by 11 year-olds in the UK, with agreement between human raters and machine marking using e-rater (the software that forms the core of *Criterion).* The results suggested there was little difference in the way more mechanical factors such as paragraphing were scored, although essays exhibiting more abstract qualities such as interest and relevance tended to be marked higher by humans.

In terms of AES effectiveness in providing formative feedback, Lee et al (2009) compared a web-based essay critiquing system developed by themselves to provide adult EFL students with immediate feedback on content and organisation for revision. A comparison was made between essays written by two groups - an experimental group receiving feedback from the web-based system and a control group who wrote their essays on the computer in the "traditional" way. There was no statistically significant difference between the two groups in essay length, or in the final scores given by two human raters.

Attali (2004) reports a large-scale study based on *Criterion*, which, as well as a holistic essay score, provides feedback on grammar, usage, mechanics, and style. 9275 essays were submitted to *Criterion* which provided feedback to the students, who then submitted a revised essay to *Criterion.* Data were analysed from the first and last (of three) essays submitted by US students in the 6th through the12th grade during the 2002-2003 school year. An overall measure of grammar, usage, mechanics, and style errors was computed by summing the individual error rates, and grammar, usage, mechanics, and style errors were counted for each essay and divided by the essay length to produce an error-rate. Results suggested that overall scores improved and essay length increased for revised submissions compared to the first submission. Similarly, organization and development scores improved and Attali (2004) claims that students were generally able to correct at least some types of error in subsequent versions of their essays.

2.3. Students' feedback preferences

Research suggests that the extent to which student-writers believe that using AES enhances their writing skills is still unknown (Lai, 2010). Denton et al (2008) compared the reactions of students to handwritten and electronic feedback using *Electronic Feedback* software. Students rated the electronic feedback superior for "markscheme clarity, feedback legibility, information on deficient aspects, and identification of those parts of the work where the student did well", and the lecturers reported taking less time to mark when using the software.

Lai (2010), however, investigated preferences among English as a foreign language (EFL) learners in Taiwan for computer-based using *MY Access* or peer feedback and found that although both forms were considered effective the learners tended to express a preference for peer feedback over computer feedback. Matsumara (2004) investigated the influence of computer-anxiety on the preferences of Japanese students for face-to-face teacher feedback, online teacher feedback, and peer feedback in EFL writing classes. The students were able to choose the kind of feedback they wished to receive, and their choices differed according to their level of computer anxiety. The essay writing of both high- and low-anxiety students improved as a result of being allowed to choose whether or not to use computers.

The relationship between student preferences and the effectiveness of different sources of feedback was investigated by Tuzi (2004), who provided 20 L2 writers with a "database-driven web site specifically designed for writing and responding", with oral feedback from friends and peers, and with feedback from face-to-face meetings with university writing centre tutors. Whilst the students tended to prefer oral feedback, there was evidence that computer-based feedback was likely to have more effect on their revision of their work than oral feedback, and encouraged them to focus on revision at the macro-rather than micro-level.

3. THE CURRENT STUDY

3.1. Background

Writing courses in the university in Egypt where the current study was conducted (Alexandria University) are intended to help trainee EFL teachers to develop their ability to produce a variety of text-types such as essays, résumés, and cover letters. Data elicited from instructors' questionnaires and interviews at the pre-treatment stage suggested that writing tasks are frequently set and that regular feedback is provided on written work. The data also suggested that instructor-feedback on students' work is intended to help students produce more, and better, writing through:

- *providing evidence of current student writing performance,*
- *identifying what knowledge and skills have been learnt,*
- *encouraging self-directed learning, and*
- *motivating students to write.*

However, interviews with instructors revealed that because of the large numbers the general practice among instructors is to select, from time to time, a small sample of written essays from those submitted during the course, and to present oral feedback on those essays to the whole class. Regular feedback is therefore provided on only a small sample of writing produced by class, and the assumption appears to be that students whose essays have not been selected will nevertheless be able to see the applicability of those comments on other students' essays to their own work. It is therefore unlikely that most students will be given any individual feedback at all on their work, and if they are given feedback, it will be given orally, and at most on a single draft. To many students, the oral feedback provided had a negative impact on their attitudes and therefore on their ability to make use of it. The aim of this study was therefore to investigate whether the use of CBF using *Criterion* would have an effect on attitudes towards feedback, and on the students' writing process and writing product.

3.2. Research questions

The study sought to answer the following research questions:
What is the effect of providing regular computer-based feedback on students' writing?

a) What evidence is there of changes in the students' attitudes?
b) What evidence is there of changes in the student' writing processes?
c) What evidence is there of changes in the student' writing products?

3.3. Sample

The pre-treatment stage involved 549 1st, 2nd, 3rd and 4th years trainee EFL teachers at a university in Egypt. Among the 549 participants, 27 participated in the interviews and 40 took part in the focus groups at the pre-writing stage. At entry level, students are similar in terms of age, previous education, and achievement at secondary school, but heterogeneous in relation to gender as females generally outnumber males in all four years of study (Students' Affairs Bureau, 2006, 2008). This stage also involved 31 instructors from the target context. The treatment and post treatment stages comprised 24 students who received CBF. These students were part of the 549 who filled in the pre-treatment questionnaire.

3.4. Research techniques and procedures

3.4.1. Pre-treatment

An initial questionnaire was carried out among students at Alexandria University, with the aim of investigating whether or not there was indeed a problem with the feedback provided on students' writing. The questionnaire was distributed to 804 students and 549 were completed and returned. 50% of the questionnaire sample came from 3rd year students (as they were the biggest class), 14% from first year, 18% from 2nd year and a similar percentage from 4th year. Follow-up interviews were carried out with a representative sample of trainees from the four years of study using *stratified and random sampling* techniques. Lists of names of all registered full-time trainees were obtained from the Students' Affairs Bureau and were then read into an Excel file and 50 names were selected at random. Letters were sent to potential interviewees and some instructors made in-class announcements when 27 agreed to be interviewed. The interviews were then followed by focus groups where some classroom visits were arranged with some instructors in the target institution and a small presentation about the overall research objectives was given. Additionally, announcements about the need for volunteers were made. A number of contact cards were made available and 40 students stepped forward and provided their contact details.

3.4.2. Treatment

Automated computer-based feedback was identified as a possible means of providing regular feedback on the participants' written work, and *Criterion* was the software selected for this study partly because it provides automated feedback on a wide range of language areas including grammar, mechanics, style, usage and content and organization. Additionally, informal evaluation by the researchers suggested that *Criterion* was relatively straightforward to use, which was especially important given the fact that the students would be expected to use the software on their own rather than in class, and that the pre-treatment questionnaire showed the level of their IT competence was average. 24 students (self-selected volunteers from the pre-treatment sample) were then recruited from among the trainee EFL teachers. A training session was held to familiarize participants with *Criterion*. They were given a total of four topics to write about at home without writing restrictions over 8 weeks, and were asked to submit an initial draft for each topic, and then to submit a revised draft using the computerized-feedback on their initial draft.

3.4.3. Post treatment

A post-treatment questionnaire was administered to all 24 subjects to identify possible impact on attitudes towards computerized-feedback, and post treatment interviews and focus groups were conducted with the same subjects. The impact on the students' writing processes and products was investigated using data recorded by the software itself as the *Criterion* web-site

records data on the written work for individual students, and the whole group. Additionally, two human raters were used to validate automated scores and feedback.

4. RESULTS

4.1. What evidence is there of changes in the trainees' attitudes?

The pre-treatment stage involved trainees' questionnaires (n549), interviews (n27) and focus groups (n40), which aimed to collect background data from trainees about instructor feedback practices and trainees' attitudes towards such practices. Data were elicited from 1st, 2nd, 3rd and 4th year students. Analysis of the data showed that, on a four-point scale (*positive*, *negative*, *neutral* and *not sure)* 25% had *negative* attitudes, 45% were neutral, 17% held *positive* attitudes and 13% were undecided.

Qualitative data from the pre-treatment interviews, and from the focus groups, confirmed a general feeling of dissatisfaction with, and distrust of, feedback practices. Because of the large numbers, oral feedback on a random sample of students' written work was adopted by instructors as the feedback strategy. As part of trainees' critical reflections in the focus groups at the pre-treatment stage, *implicitly and explicitly expressed*, on the oral feedback strategy deployed by instructors, an interesting account given was:

> I think that we all write…and we then submit…Sometimes we copy off each other and it does not matter if the original piece is good or bad because we generally depend on the fact that the instructor might not see them all, or even at all. So, it does not really make a difference whether we write or not in the first place.

Although some comments blamed the schooling system in general, the majority of trainees explicitly put the blame on *instructors*. In fact, they held instructors responsible for failing to ensure that the trainees were given regular feedback on their writing. Typical students' comments included:

> Of course it is not my responsibility- yes, the instructor has got some excuses (like large numbers), but he [instructor] should try hard even if he was under these pressures.

> I think it is the responsibility of the instructor because he is the one who plans everything and surely he knows what he wants to achieve.

Such data emphasize that trainees are not mere recipients of instructor feedback, rather they are capable of reflecting on such practice.

Having provided regular computerized-feedback on the written work of 24 students at the treatment stage, post-treatment instruments examined whether a change in the attitudes of these trainees occurred as a result. Post-treatment questionnaires, however, showed an overwhelmingly positive view towards the feedback provided by *Criterion* as compared to

instructor-feedback. The 24 trainees were asked in the questionnaire to rate their attitudes towards feedback *before* and *after* using *Criterion*. More than 88% expressed *positive* attitudes after using *Criterion*, compared to 16% before. Likewise, a noticeable reduction in *negative* attitudes was observed as only 12% of participants still had *negative* attitudes (*see figure 1*).

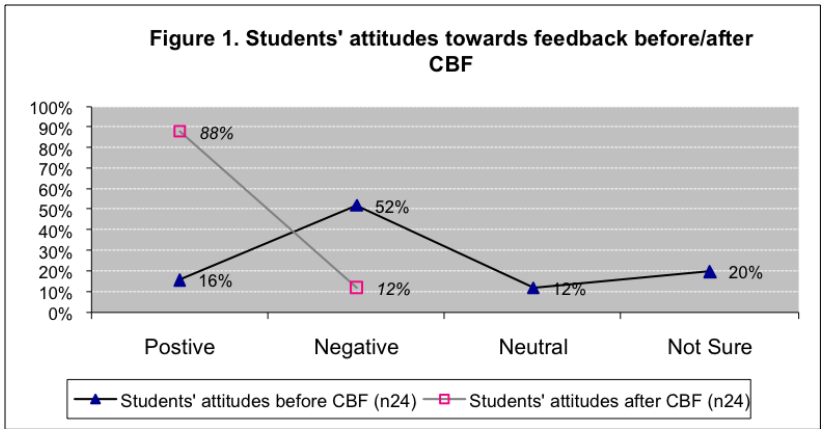

Figure 1. Students' attitudes towards feedback before/after CBF

Interestingly, the results were similarly positive in relation to trainees' views of the effect *Criterion* had on the quality of their writing. These results were confirmed in the interviews and the focus groups where participant-trainees provided further details of their awareness of the language problems *Criterion* had highlighted, and of their desire to tackle them. It was also clear from the post-treatment data that most students had noticed a change in their attitudes. Exemplary comments were:

> It [computerized-feedback] has to be used with all students because it helps and encourages me to write… it makes me eager to know the grades [holistic score]and mistakes [analytical feedback].

Qualitative data also suggested a change in attitudes towards writing. One participant commented:

> I hated writing before using Criterion because I did not know the steps and I did not know how I was evaluated…basically assessment was random…but after using Criterion I think I am much better I know my errors instantly …I develop my writing and my marks often differ from one submission to another.

So, this paper claims that the post treatment instruments suggested an overwhelming change in participant-trainees' attitudes towards feedback as well as their views about the quality of their writing.

4.2. What evidence is there of changes in the trainees' writing processes?

One of the major advantages of using technology for assessing writing is the data that can be automatically recorded, not only about trainees' final writing performances, but also about processes and procedures that might precede or accompany such performances. Providers of the *Criterion* service claim that it supports all stages involved in the writing process, i.e. *planning, writing and revising* (http://www.ets.org). Furthermore, a variety of resources are available within *Criterion* which are intended to help students at various stages in the writing process, including planning templates, spelling checkers, timers, and word counting. It was decided for the purpose of this study not to impose any restrictions on the use of these resources, in order to collect as much data as possible about the potential effect on writing processes. Participant-trainees were therefore informed about all of the available resources during the training session provided prior to actual use of *Criterion*. To investigate the impact of CBF on students' writing process, the data elicited from *Criterion* were used to examine whether or not the 24 students who received CBF planned, submitted, revised and used the feedback received on their first drafts.

4.2.1. Planning

Data from the pre-treatment stage suggested that, although pre-writing strategies are taught to trainees in the target context, the large student numbers impede instructors' attempts to monitor whether or not any of these strategies were actually being used. Qualitative data elicited from students were even inconclusive in this respect. Exemplary responses from interviews indicate that trainees know about the strategies, but they rarely use them. However, this was understandable as many students copied off each other and a considerable number never submitted their written work for feedback.

Criterion offers optional *prewriting tools* as '*assignment options'*, giving users a choice among eight different planning templates. These are intended to provide a structured approach to help students learn how to use planning strategies as they prepare to write an essay. We therefore investigated how students' generally approached their writing tasks by looking at their usage of the available planning templates, which include 1) outline, 2) list, 3) idea tree, 4) free writing, 5) idea web, 6) compare & contrast, 7) cause and effect, and 8) persuasive. Analysis of trainees' data revealed that out of 192 submissions made by participants, only *one* trainee had used a pre-writing planning template in one of the essays, and whilst follow up interviews indicated that *7* students out of 24 planned a few of their essays on paper before writing them in *Criterion*, there was no evidence that the great majority of the participants used any particular pre-writing strategy. The data therefore suggest that *Criterion* had no positive effect on the trainees' pre-writing behaviour, which supports the findings of some earlier studies (e.g. Haas, 1989).

4.2.2. Submission of written work

A general objective of the writing course in the target context is to provide exposure to and practice of different writing genres. The design of the course therefore assumes frequent writing assessment tasks requiring trainees to write and submit a number of essays during the term. Data from the pre-treatment stage on participants' perceptions of the frequency of their writing showed that, on a four-point scale only 8% of the trainees claimed to write essays '*frequently*'. Over 56% of them claimed they wrote essays '*occasionally'* or *'rarely'*, whilst 36% claimed *'never'* to write essays. Hence, conventional feedback as perceived by participants appeared to have a negative impact not only on trainees' attitudes, but also on frequency of writing and on their willingness to write. Unlike conventional writing/feedback modes, data from *Criterion* showed that the submission rate for each of the four essay topics among the 24 subjects was 192 submissions (100%). Qualitative data elicited from trainees suggested they were willing to write. The data also showed the persistence of students wanting more feedback on revised second drafts. The use of *Criterion* did, therefore, appear to motivate trainees to submit written work.

4.2.3. Revision and feedback use

Data obtained from the pre-treatment revealed that a small percentage of the subjects (about 4%) showed '*frequent'* use of the feedback given. Yet, slightly more than half of the subjects involved claimed that they only '*occasionally'* or '*rarely*' benefit from the feedback and a high percentage of respondents (45%) indicated that they '*never'* write a second draft. Qualitative data suggest that those who said that they used feedback meant in subsequent first drafts rather than in revised drafts as trainees often produced single drafts. Furthermore, many trainees showed a general feeling of lack of ownership of the feedback provided. Interestingly, one trainee stated that she rarely submitted her own essays and she further contended:

> Had he [instructor] been reading my writing regularly, I would have taken it seriously …reading it entails feedback and this is my real grade.

Additionally, the conventional feedback given was generally product-oriented where instructors provided oral group form-focused feedback on single drafts of a random sample of trainees' written work. This was the only channel through which they could see some feedback on writing, not necessarily their own, but similar in one way or another. Students were left to work out the connections between the feedback on the selected scripts and their own productions

Criterion provides analytical feedback immediately after each submission about grammar, usage, mechanics, style and content and organization, and participants were therefore expected to revise their essays and submit a second version of the same piece of writing. *Criterion* provides data for individual students showing what kind of errors were

identified in each of their submissions, but also provides aggregate data for all students across all their assignments.

Computerized-feedback appeared to show an effect on students' revision of errors in terms of grammar, usage, mechanics and style, though the nature and size of this effect varied from trainee to trainee. Perhaps most interesting was the way in which the provision of feedback appeared to at least encourage students to reflect in some detail on their writing, and on the possible problems that the computerized-feedback highlighted. Some comments were:

> I did not know before how repetition might affect writing. I discovered that repetition of words can be a mistake…especially after my feedback made this clear…I now count the words that I am skeptical about before submitting my essay.
>
> I think focusing on grammar and vocabulary is not enough. To be a good writer I need to focus on the mistakes [obtained from feedback].

This reflection probably contributed to their belief that *Criterion* helped them to study effectively without a teacher.

4.3. What evidence is there of changes in the trainees' writing products?

In examining whether or not CBF appeared to have an effect on *writing product*, the holistic scores provided by *Criterion* for each of the essays were used to see whether or not the scores show improvement between drafts and between submissions. *Criterion* extracts linguistically based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a holistic score out of 6. *Criterion* also places students into three different levels based on the scores they attain. These levels are '*doing well*', '*needs some help*' and '*needs a lot of help*'.

Results showed a consistent improvement in students' holistic scores from their first essay through to their fourth (see figures 2, 3, 4 & 5). For example, the data obtained from the students' first essay showed that a large proportion of trainees' essays (43.5%) were located in the middle of the holistic scale. According to that scale specification, 43.5% - those who obtained 3 out of 6, along with another group who scored 4 out of 6 (13%) - '*needed some help*' with their writing. Similarly, a high percentage of students (34.7%) '*needed a lot of help*' with their writing as their writing was rated 1 or 2 out of 6 on the same scale and only 8.6% of students obtained either 5 or 6 and were therefore described by *Criterion* as '*doing fine*'.

In their resubmissions, according to *Criterion*, trainees made progress in terms of both their holistic scores and the level of help they were deemed to need. It was noticed however, that in later essays, rather than correct mistakes, trainees tended to employ strategies to avoid mistakes because they were keen, as they explained in the interviews, on getting a better score. Whilst this in turn led to them making different mistakes, nevertheless by the time they

submitted the second submission for the third essay, 63.6% were in the 5 category, suggesting an overall improvement in the quality of their writing, and results for the fourth essay suggested further improvement as 63.6% scored 6, 13.6% scored 5 and yet another 13.6% scored 4. Therefore on the basis of *Criterion's* assessment of the students' work, more students were '*doing fine*', fewer students were in the '*need some help*' category, and even fewer were categorised as in '*need a lot of help*' (*see table 2*), suggesting that, in this case at least, CBF had a *positive impact on the quality of students' writing product*.

In the following figures the number of resubmissions scored for each essay varies between 21 and 23. The total number of resubmissions for each essay was, in fact, 24, i.e. all students submitted second drafts. However, for a small number of resubmissions *Criterion* was unable to award a score, and issued an "advisory", e.g. "Your essay could not be scored because some of its organizational elements could not be identified". As a result, although a total of 196 essays were submitted (24 first and 24 second submissions for each of the 4 essays), only 192 were scored by *Criterion*.

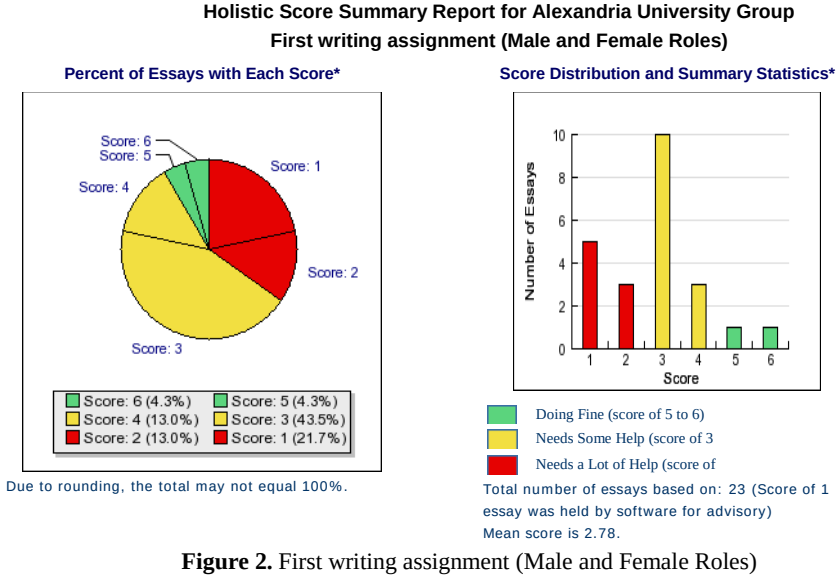

Figure 2. First writing assignment (Male and Female Roles)

Figure 3. Second writing assignment (Money and Success)

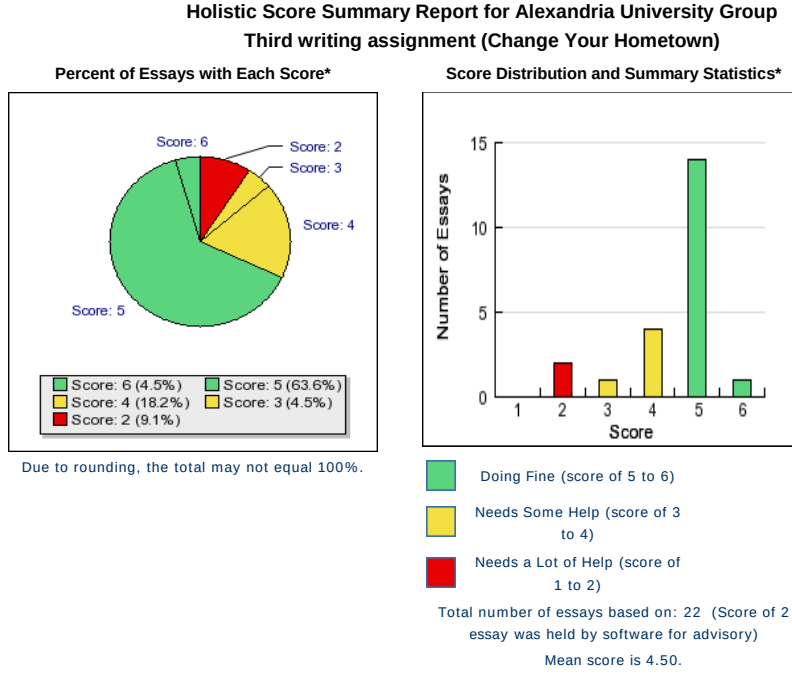

Figure 4. Third writing assignment (Change Your Hometown)

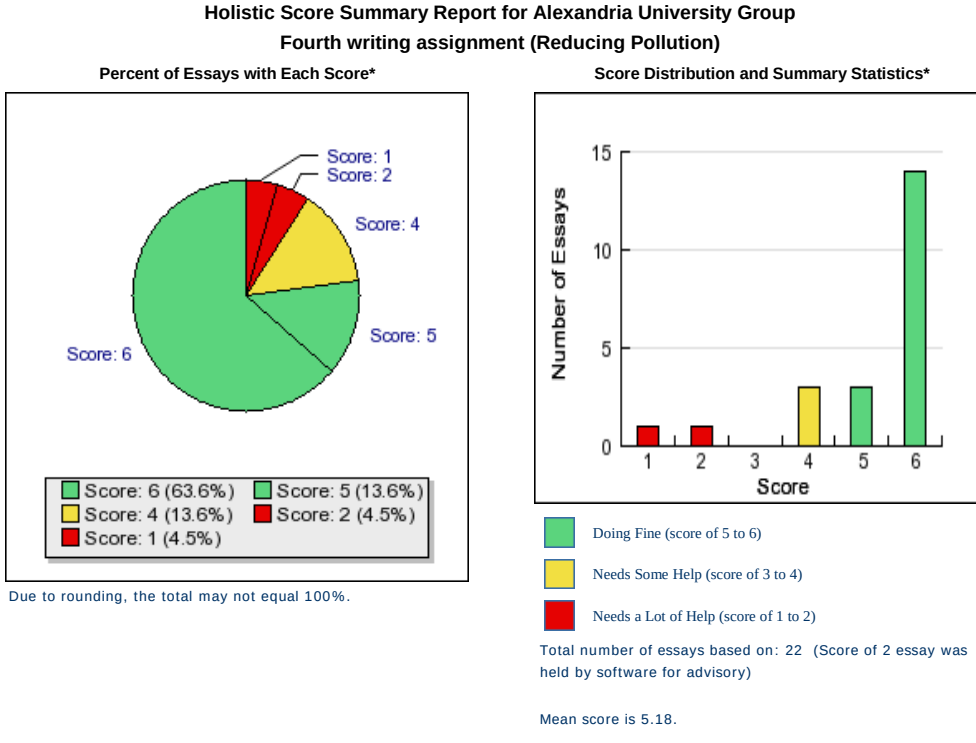

Figure 5. Fourth writing assignment (Reducing Pollution)

	1st Assignment	2nd Assignment	3rd Assignment	4th Assignment
Needed a lot of help	34.7%	4.8%	9.1%	4.5%
Needed some help	56.5%	19.1%	22.7%	18.1%
Doing fine	8.6%	76.2%	68.1%	77.2%

**Results given in the table are based on 192 submitted essays*

Table 2. *Criterion* evaluation of level of help needed by students*

Hence, the data provided by *Criterion* thus far suggest identifiable improvement in trainees' writing, or at least provide evidence that the students were making use of feedback to move closer to work that meets the criteria the software uses to evaluate their writing. The criteria used by the software may be no better nor worse than those used by instructors, but at least there is evidence of using the feedback and the criteria, whereas there is no such evidence in the case of the conventional feedback provided (or not provided) to these students in their normal writing classes.

However, in order to examine the accuracy of the feedback and the scores obtained from *Criterion*, two English language tutors who work at Newcastle University were asked to participate as human raters. The task required human raters to a) *holistically score a representative sample of trainees' writing using Criterion scoring scale, b) give primary trait feedback on the same language areas Criterion feeds students back on and c) provide placement on the same scale as that used by Criterion, i.e. doing fine, need some help and need a lot of help*. Two types of comparison were carried out in order to examine *Criterion* holistic scores on students' writing. The first one involved comparing *first* and *second* submissions of the first writing task. The holistic score awarded by *Criterion* suggested that, whilst students made progress over a number of essays, their holistic scores did not necessarily change between first and second submissions of the same essay. For this first task, therefore, 12 essays written by 6 students on the same topic (*as first and second submissions*) were selected to represent a range of holistic scores awarded by *Criterion*, and two patterns, i.e. holistic scores which improved between submissions, and scores which remained the same. Table 3 summarises the scores awarded by *Criterion* for these essays, and the corresponding scores awarded by the two human raters.

Sample	***Submission***	***Rater 1 mark***	***Rater 2 Mark***	***Criterion Mark***
Stu 1	1st submission	2	4	3
	2nd submission	3	4	4
Stu 2	1st submission	4	3	4
	2nd submission	4	3	4
Stu 3	1st submission	1	1 or 2	2
	2nd submission	5	2	3
Stu 4	1st submission	3	2 or 3	2
	2nd submission	3	2 or 3	3
Stu 5	1st submission	2	2	1
	2nd submission	1	1	1
Stu 6	1st submission	2	2	1
	2nd submission	1	1	1

* moderate inter-rater reliability between rater 1 and 2 (r =*.451*)

*moderate inter-rater reliability between first rater and *Criterion* (r =.624) and between second rater and *Criterion* (r =.499).

Table 3. Holistic scores given on 1st and 2nd submissions by human raters versus *Criterion**

Inter-rater reliability was calculated using Pearson's *r*, indicating moderate inter-rater reliability between the two human raters, and between each rater and *Criterion*.

The second comparison was between the scores awarded by *Criterion* and by the two human raters for the students' *first* submissions for their first essay, and the second

submission for their *most recent* essay. Table 4 summarises the scores awarded by *Criterion* for these essays, and the corresponding scores awarded by the two human raters.

Sample	***Submission***	***Rater 1 mark***	***Rater 2 Mark***	***Criterion Mark***
Stu 1	1st submission	2	4	3
	Recent submission	4	4	6
Stu 4	1st submission	3	2 or 3	2
	Recent submission	5	3	6
Stu 6	1st submission	2	2	1
	Recent submission	2	2	3

Table 4. Holistic Scores given on first and most recent submissions by human raters vs. *Criterion*

Inter-rater reliability calculated using Pearson's *r* showed significant inter-rater reliability between the first rater and *Criterion* (r =.839) and moderate inter-rater reliability between the second rater and *Criterion* (r =.539).

Thus, it can be claimed that the use of computerized-feedback had a positive effect - according to the data from *Criterion* - on the quality of the students' writing. However, *Criterion's* designers suggest that the system is not immune from errors, and the generally moderate level of agreement between *Criterion* holistic scores and those provided by trained professional readers suggests that the scores provided by the system should be used as just one piece of evidence about the quality of students' writing.

5. DISCUSSION AND CONCLUSION

We set out to investigate whether the use of computer-based feedback using *Criterion* would have an effect on attitudes towards feedback, writing process and product. The context of the study was a teacher training degree at Alexandria University in Egypt, where, because of large student numbers, the trainee EFL teachers were expected to produce assessed essays regularly for which they received little or no individual feedback, and no information about their marks. 31 instructors and 549 trainees took part in the pre-treatment. 24 trainees from the pre-treatment stage participated in the treatment and post treatment stages.

Evidence gathered from data recorded by *Criterion* showed that the trainees made virtually no use of the pre-writing tools that were available online, with only one participant recorded as doing any online planning (with just one essay). Follow-up interviews identified a further 7 who had planned few of their essays on paper, but on this evidence far less than half of the students appear to occasionally use pre-writing strategies. There are two issues here,

however. The first is that the fact that the majority of students did not—and therefore perhaps do not –routinely use pre-writing strategies suggests that this is an aspect of their writing skill that needs to be addressed. The second is that *Criterion* did not seem to have an effect on this aspect of their writing process, or at least the trainees did not seem to adopt what would generally be considered effective pre-writing behaviour. Why the pre-writing resources offered by *Criterion* were not used remains unclear. Perhaps greater familiarity with the tools *Criterion* provides might alter their behaviour.

As far as revision is concerned, *Criterion* was clearly effective in encouraging the students to produce second, revised versions of the essays they wrote on each of the four topics. The resubmission rate was 100%, and as data from pre-treatment suggested they virtually never produced revised versions of essays, this represented a significant change in their normal writing. In addition, the resubmissions showed evidence that the students had taken notice of the computer-based feedback as the number of errors identified in the categories used by *Criterion* generally showed a decrease in the second drafts of the essays. It would be encouraging to feel that students were not only taking notice of the feedback, but learning from it too. However, more detailed examination of the results suggested that some students were sometimes simply omitting at least some of the errors identified, rather than learning from the feedback and producing corrected versions. It was unclear whether or not this might be viewed as evidence that CBF made trainees aware of some language areas they are likely to make mistakes in. However, data from the student interviews and the focus groups identified one reason for this as a spirit of competition, i.e. the students' orientation towards getting better scores and feedback for their essays.

As for writing quality, *Criterion* not only provides feedback on particular features of the trainees' writing, but provides an overall score which is indicative of the amount of help the trainees are felt to need with their writing. These scores show an improvement over the four essays for most participants, and a fairly dramatic improvement between the first and second essays. One explanation for this may be the positive reaction on the part of the students to the feedback provided by *Criterion*, and perhaps even to the fact that they were given feedback at all. A second reason may be the avoidance strategies that some, at least, of the students developed, so the improvement may not have represented learning, or it may indicate an awareness of what they were doing wrong, but not necessarily an understanding of why, or of how they could correct some of their mistakes. It was not clear either that improved scores necessarily indicated an improvement in performance that would be acknowledged by other judges. Whilst a proper validation exercise was not one of the aims of this study, a small-scale exercise was conducted to compare marking of a selection of the same scripts by *Criterion* and two human raters. This highlighted some interesting differences between the details of the *Criterion* feedback and that given by human raters, with differing levels of consistency, and a difference in the categories of mistake that were focussed on.

Post-treatment questionnaires showed positive attitudes towards the feedback provided by *Criterion*, and were similarly positive about the effect using *Criterion* had on the quality of their writing. These results were confirmed in the focus groups and interviews.

These results suggest that the computer-based feedback on writing was effective in tackling the problems in the context in which the study was carried out. Regular and timely feedback was available to trainees, who as a result wrote essays on a regular basis, paid attention to the problems identified in their work, and revised it and produced a second draft, again on a regular basis. The quality of the writing appeared to improve, though the nature of the changes leading to that improvement merits further study, given that some students appeared to achieve better scores by using avoidance strategies. The writing processes of the trainees, and the strategies they employ would therefore benefit from further investigation, especially as the students' pre-writing strategies appear under-developed, and the availability of resources in *Criterion* to help with pre-writing remained unused. This study was also relatively short term, lasting just 8 weeks, and it is possible that the improvements identified were partly or mainly due to the novelty (or Hawthorne, or experimental), effect (McNeill and Chapman, 2005). Whether or not the improvements would be maintained would probably depend on how useful the *Criterion* feedback is in leading to better writing in a non-*Criterion* context – or at least to better college writing scores. Longer term use of computer-based feedback would certainly yield different results, as would the use of computer-based feedback with students in different contexts and at different levels of proficiency.

Perhaps most interesting from a practical point of view would be a study of a "hybrid" use of computer-based and teacher feedback. The situation in Alexandria University, where no, or virtually no, individual feedback on writing is available on a regular basis is different from many other situations where the problem for teachers is to decide how best to use the time they have to provide feedback, and in particular, what to focus their feedback on. Computer-based feedback in those contexts would probably be best used in conjunction with teacher-based feedback, and the question then is how teachers would react to sharing feedback responsibilities in such a hybrid situation, and what approaches would be most effective. Would it be preferable for teachers to rely on computer-based feedback for mechanical errors, leaving them to concentrate on the content and organisation of drafts revised on the basis of feedback provided by computer? Would it be possible to do this? We are currently working with a colleague on just such a study.

REFERENCES

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion.* Paper presented at the Paper presented at the National Council on Measurement in Education, San Diego, CA.

Black, P., and Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-47.

Black, P., and Wiliam, D. (1998). Inside the Black Box: Raising Standards through Classroom Assessment. *Phi Delta Kappan, 80*.

Bull, J., and McKenna, C. (2004). *A Blueprint for Computer-Assisted Assessment*. London: Routledge.
Chapelle, C. A., & Douglas, D. (2006). *Assessing Language through Computer Technology*. Cambridge, UK: Cambridge University Press.
Chen, C.-F., and Cheng, W.-Y. (2006, May 27, 2006). *The Use of a Computer-Based Writing Program: Facilitation or Frustration?* Paper presented at the 23 International Conference on English Teaching and Learning, Wenzao Ursuline College of Languages, Kaohsiung.
Chodorow, M., Tetreault, J., & Han, Na-Rae. (2007, 28 June, 2007). Detection of Grammatical Errors Involving Prepositions. Paper presented at the Fourth ACL-SIGSEM Workshop on Prepositions, Prague, The Czech Republic.
Chung, K., & O'Neil, H. (1997). Methodological approaches to online scoring of essays. ED 418 101: ERIC.
Coniam, D. (2009). Experimenting with a computer essay-scoring program based on ESL student writing scripts. *ReCALL, 21*, 259-279.
Denton, P., Madden, J., Roberts, M., & Rowe, P. (2008). Students' response to traditional and computer-assisted formative feedback: A comparative case study. *British Journal of Educational Technology, 39*(3), 486-500.
Dikli, S. (2006). An Overview of Automated Scoring of Essays. *The Journal of Technology, Learning, and Assessment (J.T.L.A), 5*(1), 1-36.
Dmytrenko-Ahrabian, M. O. (2008). Criterion Online Writing Evaluation Service Case Study: Enhancing faculty attention and guidance.
Ellison, R. (2007). *Criterion Online Writing Evaluation Service Case Study: Writing benchmarks for every student*. Retrieved 11.08, 2009, from http://www.ets.org/Media/Products/Criterion/pdf/5796_EastTexas.pdf
Ferris, D. (2003). *Response to Student Writing: Implications for second language students*. Mahwah, New Jersey: Lawrence Erlbaum Associates.
Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports student learning. *Learning and Teaching in Higher Education, 1*(1), 3-31.
Grimes, D., & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *JTLA, 8*(6), 1-43.
Hamp-Lyons, L. (2001). Fourth generation writing assessment. In Silva, T. and Matsuda, P (Ed.), *On second language writing*. Mahwah: Lawrence Erlbaum Associates.
Hughes, A. (2003). *Testing for language teachers*. Cambridge, UK: Cambridge University Press.
Hutchison, D. (2007). An evaluation of computerised essay marking for national curriculum assessment in the UK for 11-year-olds. *British Journal of Educational Technology, 38*(6), 977-989.
Huxham, M. (2007). Fast and effective feedback: are model answers the answer? *Assessment & Evaluation in Higher Education, 32*(6), 601-611.
Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing: State of the Art. *Language Teaching, 39*(2), 83-101.
Jacobs, G., Curtis, A., Brain, G. and Huang, S. (1998). Feedback on student writing: taking the middle path. *Journal of Second Language Writing, 7*(3), 307-317.
Konza, R. B & Johnston, J. (1991). The technological revolution comes to the classroom. *Change, 23*(1), 10-23.
Lai, Y.-h. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology, 41*(3), 432-454.
Lee, I. (2007). Feedback in Hong Kong secondary writing classrooms: Assessment for learning or assessment of learning? *Assessing Writing, 12*(3), 180-198.
Lee, C., Wong, K., Cheung, W., & Lee, F. (2009). Web-based essay critiquing system and EFL students' writing: A quantitative and qualitative investigation. *Computer Assisted Language Learning, 22*, 57-72.
Matsumura, S., & Hann, G. (2004). Computer Anxiety and Students' Preferred Feedback Methods in EFL Writing. *The Modern Language Journal, 88*(iii), 403-415.
McNeill, P. and Chapman, S. (2005) *Research Methods.* Taylor & Francis Ltd.

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

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated essay scoring* (No. GRE® Board Professional Rep. No. 98-08bP, ETS RR-01-03). NJ: ETS.

Rudner, L., & Gagne, P. (2001). An overview of three approaches to scoring written essays by computer. *Practical Assessment, Research & Evaluation, 7*(26), 1-6.

Rudner, L., and Liang, T. (2002). Automated Essay Scoring Using Bayes' Theorem. *The Journal of Technology, Learning, and Assessment (J.T.L.A), 1*(2), 1-22.

Shermis, M., and Burstein, J. (2003). *Automated Essay Scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.

Shermis, M., Raymat, M. V., & Barrera, F. (2003). *Assessing writing through the curriculum with Automated Essay Scoring* (No. ED 477 929): ERIC.

Tuzi, F. (2004). The impact of e-feedback on the revisions of L2 writers in an academic writing course. *Computers and Composition, 21*(2), 217-235.

Ussery, R. (2007). *Criterion Online Writing Evaluation Case Study: Improving writing skills across the curriculum*. Retrieved 11 August, 2009, from http://www.ets.org/Media/Products/Criterion/pdf/4216_NCarol3.pdf

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.

APPENDIX 1

Sample questions from the pre-treatment students' questionnaire

Question 17**. How would you describe YOUR attitude towards instructor-feedback in the writing course?**

a. Positive b. Negative c. Neutral d. Not Sure

Question 7. With what material do you prefer to write your essays?

a. Paper and pencil b. the Computer c. Both

Question 10. How often does <u>YOUR INSTRUCTOR</u> do the following?

	Frequently	Occasionally	Rarely	Never
e. Assign writing tasks	1	2		3
f. Mark essays	1	2		3
g. Give essays back with feedback	1	2		3
h. Give you detailed comment on your writing	1	2		3
i. Give you a brief comment on your writing	1	2		3
j. Ask you questions about your writing	1	2		3

Question 11. How often do <u>YOU</u> do the following?

	Frequently	Occasionally	Rarely	Never
a. Write an essay	1	2		3
b. Ask your instructors questions about your writing in class	1	2		3
c. Ask your instructors questions about your writing off class	1	2		3
d. Make use of the feedback on your essays	1	2		3
e. Use the computer to write your essays	1	2		3
g. Use the computer for reasons other than writing	1	2		3

Sample questions from the Post CBF Questionnaire

What is your opinion about the use of *Criterion* in the assessment of writing? *(you can write in Arabic)*

Explain in details how the use of computerized-feedback <u>managed</u> to or <u>failed</u> to help you?

How would you describe your attitudes towards feedback <u>BEFORE</u> the use of computers?

How would you describe your attitudes towards feedback <u>AFTER</u> the use of computers?

