Chapter XLVIII
Automated Essay Scoring Systems

Dougal Hutchison
National Foundation for Educational Research, UK

ABSTRACT

This chapter gives a relatively non-technical introduction to computer programs for marking of essays, generally known as Automated Essay Scoring (AES) systems. It identifies four stages in the process, which may be distinguished as training, summarising mechanical and structural aspects, describing content, and scoring, and describes how these are carried out in a number of commercially available programs. It considers how the validity of the process may be assessed, and reviews some of the evidence on how successful they are. It also discusses some of the ways in which they may fall down and describes some research investigating this. The chapter concludes with a discussion of possible future developments, and offers a number of searching questions for administrators considering the possibility of introducing AES in their own schools

I. INTRODUCTION

Constructed response material is being increasingly used as part of the overall assessment process, as awareness of the limitations of multiple choice testing becomes more widespread. Such constructed material- portfolios, project work, essays- while not completely supplanting closed response materials, offers the possibility of assessing a wider, and possibly more valid, range of skills.

Essays in particular require the ability to construct, organise and justify ideas, all essential skills in many jobs today. Further, practice in writing essays is in itself an important part of developing writing and communication skills. From the other side of the desk, teachers find essay marking an important part of the formative assessment process, which can give important insights into the extent to which their pupils have grasped a topic, and are able to communicate their understanding. The downside of this of course is the cost in time

and money of doing this. In England and Wales the marking of the written section of the National Curriculum Assessment cost some £18 million in a single year (Gunter, 2004), and it has been stated (Mason and Grove-Stephenson, 2002) that teachers take up to 30 per cent of their time in marking. There is clearly an opportunity for some kind of automation of the procedure.

Computers, and applications which make use of their specialist properties, it seems, are everywhere in our lives, from our correspondence to our toasters. They seem similarly ubiquitous in education settings, potentially delivering the subject, assessing, and interpreting the results. In some instances, too, they help blur the distinction between the learning aspect of the process, and the assessment (Boyle and Hutchison, 2008). Thus for example, a flight simulator could be used in the first stages of training of pilots, and at the same time give an impression of when it might be safe to let the trainee loose on an actual aeroplane. Assessments can be delivered by computer: for example lab work can be assessed without the accompanying spillage of water. Some types of testing would be extremely difficult to score without computer input, for example any kind of IRT scoring, and the process of delivery and assessment is combined in computer adaptive testing, for example in the NFER-Mental Mathematics test (Vappula, Morrison, Hutchison and Boyle, 2004). However, one aspect of assessment that has not been extensively affected by computerisation, until relatively recently, is that of essay marking. Yet this too is changing.

II. AUTOMATED ESSAY SCORING SYSTEMS

A number of programs (Automated Essay Scoring, or AES, Systems) are described in the literature. These vary in the extent to which they are generally available in a usable form: this chapter will concentrate for the most part on the larger, commercially available systems. The first to be developed, by some margin, is Project Essay Grade (PEG), which first started in the 1960s, but has more recently been revived in an updated form (Page, 2003). Three other major systems are Intellimetric ™ (Elliot, 2003) and the Intelligent Essay Assessor ™ (IEA) (Landauer, Laham and Foltz, 2003) and *e-rater* ® **(**Burstein, 2003, Attali and Burstein, 2006). The work of Larkey and Croft (2003) has been significant in this area, and while it appears that there is no corresponding commercially available program, the program BETSY ((Rudner and Liang, 2002)), currently freeware, appears to share much of the same theoretical features.

There is a bewildering and impressive range of techniques used in such AES systems, and in what follows the functioning of a few will be described as far as possible. The programs differ in the extent to which it is possible to discern what they actually do. Some, such as *e-rater*®, do describe publicly and quite extensively what they do: others cloak what they do under a cloak of commercial confidentiality. The descriptions to follow are taken from those produced by the developers, rather than from any personal knowledge of the inside workings of the packages. These systems are under constant development, so aspects of these descriptions may well date: for the latest information consult the respective websites. The majority of the systems I have managed to identify are centred round English, though Intellimetric is also operational in other languages, Malay, Turkish, Spanish and Chinese (Edelblut, 2008). Burstein and Chodorow (1999) discuss AES for non-native English speakers.

Much of what follows will refer to *e-rater*®, because of the quality of its documentation, but the absence of reference to other AES packages should not necessarily be taken as indicating they lack the relevant features, merely that these are not publicly documented.

Typically AES systems have four different aspects, though some may differ in details.