Chapter 23
Automated Scoring in Assessment Systems

Michael B. Bunch
Measurement Incorporated, USA

David Vaughn
Measurement Incorporated, USA

Shayne Miel
LightSide Labs, USA

ABSTRACT

Automated scoring of essays is founded upon the pioneer work of Dr. Ellis B. Page. His creation of Project Essay Grade (PEG) sparked the growth of a field that now includes universities and major corporations whose computer programs are capable of analyzing not only essays but short-answer responses to content-based questions. This chapter provides a brief history of automated scoring, describes in general terms how the programs work, outlines some of the current uses as well as challenges, and offers a glimpse of the future of automated scoring.

INTRODUCTION

Automated scoring of written essays and other types of student responses has been a dream of educators, educational assessment specialists, and computer scientists for over 50 years. Educators have looked for ways to teach more efficiently; assessment specialists have looked for objective ways to produce scores for examination types that have historically been considered quite subjective; and computer scientists have viewed automated scoring as a field ripe for exploiting the power of the computer. In this chapter, we explore the history of automated scoring and examine the development of systems before turning our attention to current applications and challenges. We conclude with a look to the future.

A BRIEF HISTORY OF AUTOMATED SCORING

Automated scoring has a rich history, dating back to *Natural Language and the Computer,* edited

DOI: 10.4018/978-1-4666-9441-5.ch023

by Paul Garvin (1963). That book contained an overview and 16 essays on various aspects of solving natural language problems with high-speed computers. The tone of the book, as well as a clue to its application to automated essay scoring, is clearly expressed in the chapter by L. C. Ray (1963, p. 95):

These new tools are important in research because they promise significant economies, especially in terms of time, in operations involving massive paperwork. They are equally important in that they can be utilized to carry out tasks that are not now being done because other means cannot accomplish the job or cannot do it in time for the results to be of use.

Garvin's book was an outgrowth of the artificial intelligence (AI) movement sparked by British mathematician (and famed wartime codebreaker) Alan Turing. In the years following World War II, Turing and others turned their attention from the narrow task of codebreaking to the more general application of artificial intelligence to a host of problems. The transition from decoding secret messages to deconstructing and reconstructing prose was a natural one.

Although the term "automated essay scoring," or AES, did not appear formally in the research lexicon until Shermis & Burstein's 2003 publication of *Automated Essay Scoring: A Cross-Disciplinary Perspective*, the computer scoring of student essays traces its origins to the pioneering work of Ellis Batten Page (1924–2005). Page, widely acknowledged as the father of automated essay scoring, was a pioneer in the application of the computer to the scoring of student essays. His focus was specifically on writing quality, as opposed to correctness of content (e.g., the communicative effectiveness of a five-paragraph essay as opposed to the historical accuracy of an exposition on the Treaty of Ghent). Page (1966) reported on an early effort to understand how human graders applied evaluation criteria to student essays and to recreate those criteria in a computer program. That program, Project Essay Grade, or PEG®, was designed to score student essays using mainframe computers in the 1960s.

Dr. Page and his colleagues coined two new terms: *trin* and *prox*. A trin is an intrinsic characteristic of writing, such as diction or style. A prox is a quantifiable approximation of that intrinsic characteristic. For example, a prox for diction might be the proportion of words in a fifth grader's essay found on word lists for sixth grade and above. A prox for style might be the number of times the word "because" appears in an essay, as such words are proxies for complex sentences with subordinate clauses. These terms have since been replaced by "features," and there is no practical distinction between intrinsic and objectified features.

In their initial experiment, the PEG team had four human judges (English teachers) read 276 essays written by high school students. Dr. Page and his colleagues then translated several trins into 31 proxes —predictors of overall essay quality. All essays were keypunched, and PEG calculated their scores on the 31 proxes and then correlated those scores with the overall quality scores assigned by human judges. The multiple R for the initial run was .71, with shrinkage to .65. PEG had produced total score estimates that correlated with scores assigned by humans about as well as individual human graders' scores correlated with one another. Documenting PEG's early and somewhat startling results, Page wrote, "The results of our first data run, read at 11 p.m. one night last May, were truly stunning, so much so that my colleagues and I spent the next hours in champagne and excited talk" (Page, 1966, p. 241).

Although the experiment was a huge success, it left the team with some questions.

1. What about input? Who is going to transcribe all those handwritten essays?
2. What about the gifted student who is offbeat and original?