# Chapter 3
# Automated Essay Scoring Using Deep Learning Algorithms

**Jinnie Shin**
*University of Alberta, Canada*

**Qi Guo**
*Medical Council of Canada, Canada*

**Mark J. Gierl**
*University of Alberta, Canada*

## ABSTRACT

*The recent transition from paper to digitally based assessment has brought many positive changes in educational testing. For example, many high-stakes exams have started implementing essay-type questions because they allow students to creatively express their understanding with their own words. To reduce the burden of scoring these items, the implementation of automated essay scoring (AES) systems have gained more attention. However, despite some of the successful demonstrations, AES still encountered many criticisms from practitioners. Such concerns often include prediction accuracy and interpretability of the scoring algorithms. Hence, overcoming these challenges is critical for AES to be widely adopted in the field. The purpose of this chapter is to introduce deep learning AES models and to describe how certain aspects of the models can be used to overcome the challenges of prediction accuracy and interpretability of the scoring algorithms.*

## INTRODUCTION

High-stakes testing is in the process of transitioning from paper to computer-based assessment. While the initial transitions have focused primarily on administrative benefits, such as increased test security and flexible testing schedules, the more recent transitions are focused on the use of new item formats. For example, the National Assessment of Educational Progress (NAEP) introduced innovative item types—such as interactive scenario-based questions—with their new digital-based assessment environment.

DOI: 10.4018/978-1-7998-3476-2.ch003

The purpose of these new item types is to provide more authentic assessment opportunities for students (NAEP, 2018). Such items often require students to express their understanding in a creative way using their own words, thereby, invoking higher-order reasoning and complex thinking skills (Scully, 2017).

With traditional paper-based assessment, selected-response items (e.g., multiple-choice questions) are often used because they are efficient to administer, they are easy to score objectively, and they can be used to sample a wide range of content domains in a relatively short time using a single test administration (Haladyna & Rodriguez, 2013; Rodriguez, 2016). Compared to essays and other written-response tasks, which are prone to subjective scoring and which require more time for recording answers, selected-response questions can be scored more accurately and they require students to spend less time recording answers.

However, written-response items do have many benefits. They provide evidence of students' composition and organization skills, grammatical knowledge, background knowledge, and analytic thinking and reasoning skills. Therefore, to promote the use of written-response tasks that can be used to evaluate student understanding in a creative and less restrictive way, overcoming such disadvantages stemming from scoring and administration procedures is critical in the digitally-based assessment.

Automated essay scoring (AES) was first developed to help overcome this scoring and administration problems by encouraging cost- and time-efficient marking procedures of written-response questions (Page, 1967). Traditional scoring procedures often consist of a minimum of three scorers to ensure scoring reliability and the fundamental idea of AES was to introduce a system to replace the third marker thereby saving time and money. Thus, the machine-replaced third marker can be trained based on how the scoring and grading was made by the other two human markers. To do so, the AES system has to identify deterministic linguistic features that human raters used to identify essay quality. Such features often include the length of essays, number of words, word usage, and sentence complexity. Then, the AES system attempts to learn a scoring pattern or a rule close to the human raters' using those features. When successfully implemented, AES can speed up the scoring process significantly. Moreover, it can bring several surprising benefits, such as improving the consistency of scoring and the possibility of providing instant feedback to students on their performance (Gierl, Latifi, Lai, Boulais, & De Champlain, 2014).

However, despite these potential benefits, traditional AES encountered many challenges that must be overcome before it is widely adopted in the field. Such concerns commonly stem from difficulties in selecting appropriate features and establishing the interpretability of the scoring systems (Zaidi, 2016). Selecting deterministic features directly connected to essay quality is laborious and requires tremendous amounts of linguistic knowledge. Moreover, as many commercial vendors concealed the information about features as proprietary information, AES frameworks are not accessible for most practitioners. In addition, the 'black box' nature of the traditional AES systems could not provide clear information regarding how the system arrives at the final scoring decision. Therefore, proper validation could not be made by human markers about the machine's scoring algorithms.

One of the promising directions in AES that may help address these drawbacks is the application of deep learning algorithms. Intuitively, deep learning aims to achieve more abstract and comprehensive associations among the features to achieve improved performance (Simpson, 2015). To do so, the algorithms connect input—student essays—and output— essay scores—using several "deep" layers compared to traditional machine learning. With the capacity provided by a deeper structure, these learning models can decide which features to extract without human intervention and learn comprehensive rules to make good score predictions. Moreover, because these algorithms are capable of learning features automatically in an end-to-end manner, it does not require extensive knowledge in linguistics to determine which features to include in the model (Williams & Zipser, 1989). Many studies conducted in the last decade