Digital Module 18: Automated Scoring

Sue Lottridge, Cambium Assessment
Amy Burkhardt, University of Colorado, Boulder
Michelle Boyer, Center for Assessment

Module Overview

In this digital ITEMS module, Dr. Sue Lottridge, Amy Burkhardt, and Dr. Michelle Boyer provide an overview of *automated scoring.* Automated scoring is the use of computer algorithms to score unconstrained open-ended test items by mimicking human scoring. The use of automated scoring is increasing in educational assessment programs because it allows scores to be returned faster at lower cost. In the module, they discuss automated scoring from a number of perspectives. First, they discuss benefits and weakness of automated scoring, and what psychometricians should know about automated scoring. Next, they describe the overall process of automated scoring, moving from data collection to engine training to operational scoring. Then, they describe how automated scoring systems work, including the basic functions around score prediction as well as other flagging methods. Finally, they conclude with a discussion of the specific validity demands around automated scoring and how they align with the larger validity demands around test scores. Two data activities are provided. The first is an interactive activity that allows the user to train and evaluate a simple automated scoring engine. The second is a worked example that examines the impact of rater error on test scores. The digital module contains a link to an interactive web application as well as its R-Shiny code, diagnostic quiz questions, activities, curated resources, and a glossary.

Key words: automated scoring, hand-scoring, machine learning, natural language processes, constructed response items

---

Prerequisite Knowledge

The intended audience is psychometricians who range in experience and areas of specialization and know very little about automated scoring. To get the most out of the module, it might be beneficial to have a basic understanding of the areas below to fully appreciate how automated scoring works and how it fits into the psychometric workflow:

- Industry standards for validity and reliability (e.g., as described in the AERA, APA, & NCME standards)
- Foundational statistical concepts (e.g., means, variation, correlations)
- Development processes for open-ended test items
- Human scoring processes (e.g., rubric development) and quality-control metrics (e.g., reader agreement indices)

Reading the following NCME ITEMS print modules may serve as a useful introduction to the prerequisite knowledge:

- Module 1: Performance Assessments: Design and Development
- Module 15: Assessing Student Achievement with Term Papers and Written Reports
- Module 3: Reliability of Scores from Teacher-Made Tests
- Module 43: Data Mining for Classification and Regression
- Module 42: Simulation Studies in Psychometrics

These modules and others are available for free in the ITEMS portal.

Learning Objectives

Upon completion of this module, learners should be able to:

- Understand how their analytic work might intersect with automated scoring
- Know that there are different scoring sources (i.e., human, machine)
- Identify the decision points around automated scoring that psychometricians make
- Understand how automated scores are produced
- Know the key dimensions of an automated scoring engine validation program
- Examine the impact of rater error on common agreement metrics

---

Module Structure

The digital module is divided into the following sections, which can be reviewed sequentially or independently [*approximate completion times in parentheses*].

- Module Introduction [*5 Minutes*]
- Section 1: Conceptual Foundations [*25 Minutes*]
- Section 2: Overview of Automated Scoring Processes [*25 Minutes*]
- Section 3: How Automated Scoring Works [*25 Minutes*]
- Section 4: Developing a Program of Validation [*25 Minutes*]
- Section 5: Data Activity [*20 Minutes*]
- Section 6: Worked Example [*15 Minutes*]
- Section 7: Quizzes [*20 Minutes*]

In the portal site, there is a video version of the core content as well as a handout with all core slides along with other materials.

Instructors

Susan Lottridge, *Cambium Assessment, Inc.*

Sue Lottridge, Ph.D. is a Senior Director of Automated Scoring at the Cambium Assessment, Inc. (CAI). In this role, she leads CAI's machine learning and scoring team on the research, development, and operation of CAI's automated scoring software. This software includes automated essay scoring, short answer scoring, automated speech scoring, and an engine that detects disturbing content in student responses. Dr. Lottridge has worked in automated scoring for twelve years and has contributed to the design, research, and use of multiple automated scoring engines including equation scoring, essay scoring, short answer scoring, alert detection, and dialogue systems. She earned her Ph.D. from James Madison University in assessment and measurement (2006) and holds Masters' degrees in Mathematics and in Computer Science from the University of Wisconsin-Madison (1997).

Amy Burkhardt, *University of Colorado – Boulder*

Amy Burkhardt is a PhD Candidate in Research and Evaluation Methodology with an emphasis in Human Language Technology at the University of Colorado, Boulder. She has been involved in the development of two automated scoring systems. Ongoing research projects include the automatic detection of students reporting harm within online tests, the use of machine learning to explore public discourse around educational policies, and considerations in psychometric modeling when making diagnostic inferences aligned to a learning progression.

Michelle Boyer, *Center for Assessment*

Michelle Boyer, Ph.D. is a Senior Associate at The National Center for the Improvement of Educational Assessment, Inc. Dr. Boyer consults with states and organizations on such issues as assessment systems, validity of score interpretations, scoring design and evaluation criteria for both human and automated scoring, assessment literacy, and score comparability. She is also a regular contributor to professional publications and the annual conferences of AERA, NCME, and CCSSO. Her most recent research focuses on evaluating the quality of automated scoring and its impact test score scales and test equating solutions. Dr. Boyer earned her Ph.D. from the University of Massachusetts, Amherst in Research, Educational Measurement, and Psychometrics (2018).

Instructional Designer

André A. Rupp, *Mindful Measurement*

André is the co-author and co-editor of two award-winning interdisciplinary books entitled *Diagnostic Measurement: Theory, Methods, and Applications* (2010) and *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (2016) and has just published the *Handbook of Automated Scoring: Theory into Practice* (2020). His research synthesis- and framework-oriented work has appeared in a wide variety of prestigious peer-reviewed journals. Among other things, he is passionate about improving processes for interdisciplinary collaborations during the development and implementation of scoring solutions for digitally-delivered assessments. Consequently, he is very excited to serve as the associate editor / lead instructional designer of the ITEMS portal for NCME whose mission is to provide free digital resources to support self-directed learning and professional development.