Chapter 24
Automated Scoring of Multicomponent Tasks

William Lorié
Questar Assessment, Inc., USA

ABSTRACT

Assessment of real-world skills increasingly requires efficient scoring of non-routine test items. This chapter addresses the scoring and psychometric treatment of a broad class of automatically-scorable complex assessment tasks allowing a definite set of responses orderable by quality. These multicomponent tasks are described and proposals are advanced on how to score them so that they support capturing gradations of performance quality. The resulting response evaluation functions are assessed empirically against alternatives using data from a pilot of technology-enhanced items (TEIs) administered to a sample of high school students in one U.S. state. Results support scoring frameworks leveraging the full potential of multicomponent tasks for providing evidence of partial knowledge, understanding, or skill.

INTRODUCTION

Assessment of real-world skills increasingly requires development and efficient scoring of test items that go beyond standard multiple-choice and open-response formats. There are long-acknowledged limitations on what can be tested with the former. Developing, field testing, administering, and scoring the more interesting open-ended items involves costly training and marking, even when some or all of the scoring is eventually handled through the application of machine learning algorithms. Open-ended test items have also been criticized – rightly or not – for their "subjectivity" and the inter-rater variance they introduce and which contributes to test score unreliability.

DOI: 10.4018/978-1-4666-9441-5.ch024

An assessment middle ground has emerged which promises a higher level of (at least) face validity while at the same time avoiding the costs and criticisms associated with open-ended items. This chapter addresses that middle ground, which consists of test items with constrained responses that can be enumerated and ordered a priori in terms of quality. Thus, they can be scored automatically.

Interaction-based problems in the Problem Solving exam from the Organisation for Economic Cooperation and Development (OECD) Programme for International Student Assessment (PISA), simulations in the American Institute of CPA (AICPA) Uniform CPA Examination, vignettes in the National Council of Architectural Registration Boards (NCARB) Architect Regis-

tration Examination, and multi-select and other new item types in the National Council of State Boards of Nursing (NCSBN) National Council Licensure Examination (NCLEX) are just a handful of examples of these kinds of assessment items in real-world skills exams.

The extent to which these tasks are superior to either multiple-choice or open-response formats for real-world skills assessment is a critical question beyond the scope of this chapter. The answer depends on the degree to which salient aspects of the environments in which the target skills are exhibited can be modeled in terms of constrained choices among enumerable alternatives. Many real-world skills seem to have these features: There are a limited number of actions that can be taken by pilots to respond to a particular set of readings on an airplane dashboard, by nurses in response to given specific symptoms presented by a patient, and by hotel reservation agents when a request is made for a block of rooms with special requirements for specific individuals. Contexts like these are most amenable to assessment through the types of items discussed here.

This chapter presents a framework for scoring these types of items, or parts of them. The framework examines the relationship between the response space of a test item (that is, all of the different ways in which one may respond to the item) and the evaluation function that transforms that space to yield a qualitative ordering of responses from worst to best. This general, abstract approach extends beyond what is required for standard multiple choice test items, simple open-ended items, and most other tasks the responses to which are evaluated through human scoring or machine learning algorithms that mimic human scoring. For these more typical tasks, either (a) the response space is small and the evaluation function is binary (as with multiple choice items), (b) the response space is large but most of it can be ignored (as with items requiring the examinee to supply a word, a number, or the like), or (c) the evaluation function is non-deterministic but rendered more reliable through training (as with human or machine scoring of essays).

In contrast, the item formats discussed here feature expanded (but finite) response spaces that potentially carry a great deal of information about response quality – an important reason test developers make use of them to begin with. Introducing human judgments into the evaluation functions for these items is not optimal – not only because of the possibility of introducing inconsistencies in the scoring of identical responses, but also because of the time and expense associated with collecting and modeling such judgments.

Thus, a different approach is needed. This chapter introduces the general concept of a multicomponent task, and illustrates how several recently popular test item types, such as "drag-and-drop," "hot spot," and "select all that apply," are simple examples of multicomponent tasks. These simple tasks form the foundation for more complex tasks that are, in turn, featured in assessments of real-world skills.

Essential topics for understanding multicomponent tasks include the maximum and optimal sizes of the evaluation space (in other words, how many points the item can or should support), logical dependencies between components, the role of the default response state, and the information content of real responses to multicomponent items.

This chapter takes as its point of departure a psychometric perspective in covering these topics, providing background on attempts to model responses to related tasks within the framework of item response theory. After distinguishing multicomponent tasks from other types of complex item formats, principles for an effective response coding schema for various simple multicomponent tasks are proposed. Task types, their parameters, and response spaces are described. Response evaluation functions for one type of multicomponent task are proposed and their relative merits discussed. Three alternative scoring criteria – two of them