Measurement 114 (2018) 16–24

Contents lists available at ScienceDirect

Measurement

journal homepage: www.elsevier.com/locate/measurement

Exploring the item features of a science assessment with complex tasks

Tina Collier*, Linda Morell, Mark Wilson

Berkeley Evaluation and Assessment Research (BEAR) Center, University of California, Berkeley, 2000 Center Street, Suite 301, Berkeley, CA 94704, USA

ARTICLE INFO

Keywords:
Linear logistic test model
Item explanatory models
Item features

ABSTRACT

Item explanatory models have the potential to provide insight into why certain items are easier or more difficult than others. Through the selection of pertinent item features, one can gather validity evidence for the assessment if construct-related item characteristics are chosen. This is especially important when designing assessment tasks that address new standards. Using data from the Learning Progressions in Middle School Science Instruction and Assessment (LPS) project, this paper adopts an "item explanatory" approach and investigates whether certain item features can explain differences in item difficulties by applying an extension of the linear logistic test model. Specifically, this paper explores the effects of five features on item difficulty: type (argumentation, content, embedded content), scenario-based context, format (multiple-choice or open-ended), graphics, and academic vocabulary. Interactions between some of these features were also investigated. With the exception of context, all features had a statistically significant effect on difficulty.

1. Introduction

This paper adopts an "item explanatory" approach, where the focus is on investigating whether certain item features can explain differences in item difficulties by applying an extension of the linear logistic test model (LLTM; [6] to a middle school science assessment that was designed to follow the Next Generation Science Standards (NGSS; [15]. Explanatory item response models (EIRMs; [5] have the potential to provide explanations for the item responses, unlike descriptive models where item responses are merely described by the estimated parameters. While more traditional models output a list of estimated item difficulties, an "item explanatory" approach results in a list of estimated difficulties for item features. These item features must be selected a priori and, if content-related features are chosen, have the potential to provide strong content validity support for an assessment. Specifically, this paper explores the effects of five features on item difficulty: type, context, format, graphics, and academic vocabulary.

1.1. The Next Generation Science Standards (NGSS)

The NGSS is a U.S. initiative designed to increase understanding of science, create common standards for teaching across the U.S., and develop more interest in science in school-age students in the hopes that more of them will major in a science-related area of study in college. The NGSS provides performance expectations to reflect a reform in science education that includes three dimensions: (1) developing disciplinary core ideas (DCI), (2) linking these core ideas across disciplines or crosscutting concepts, and (3) engaging students in scientific and engineering practices—based on contemporary ideas about what scientists and engineers do. The emphasis, in particular, is on integrating these three dimensions so that core ideas are not taught in isolation, but connect to larger ideas that also involve real-world applications. Rather than learn a wide breadth of disconnected content topics, the goal is to develop a deeper understanding of a few core ideas that set a strong foundation for all students after high school. The Learning Progressions in Middle School Science Instruction and Assessment (LPS) project, described in the next section, examined two of these three dimensions and designed an assessment to reflect their integration.

1.2. Item features for the Learning Progressions in Middle School Science Instruction and Assessment (LPS) project

One of the main research goals for the Learning Progressions in Middle School Science Instruction and Assessment (LPS) project,[1] was to explore the relationship between science content knowledge, a DCI, and scientific argumentation, a scientific practice. To further explore this relationship, the assessment was divided into three "complex tasks", which consist of three item *types*: (1) argumentation items assessing argumentation competency in a specific scientific context (e.g. two students arguing over what happens to gas particles placed in a container), (2) content science items embedded within the same

* Corresponding author at: 211 East Chicago Avenue, Chicago, IL 60611, USA.
E-mail address: colliert@ada.org (T. Collier).

[1] Details of the project are not discussed here, though several resources are available for interested readers [16–18,22,26,27].

http://dx.doi.org/10.1016/j.measurement.2017.08.039
Received 5 April 2017; Received in revised form 9 June 2017; Accepted 28 August 2017
Available online 31 August 2017
0263-2241/

Measurement 114 (2018) 16–24

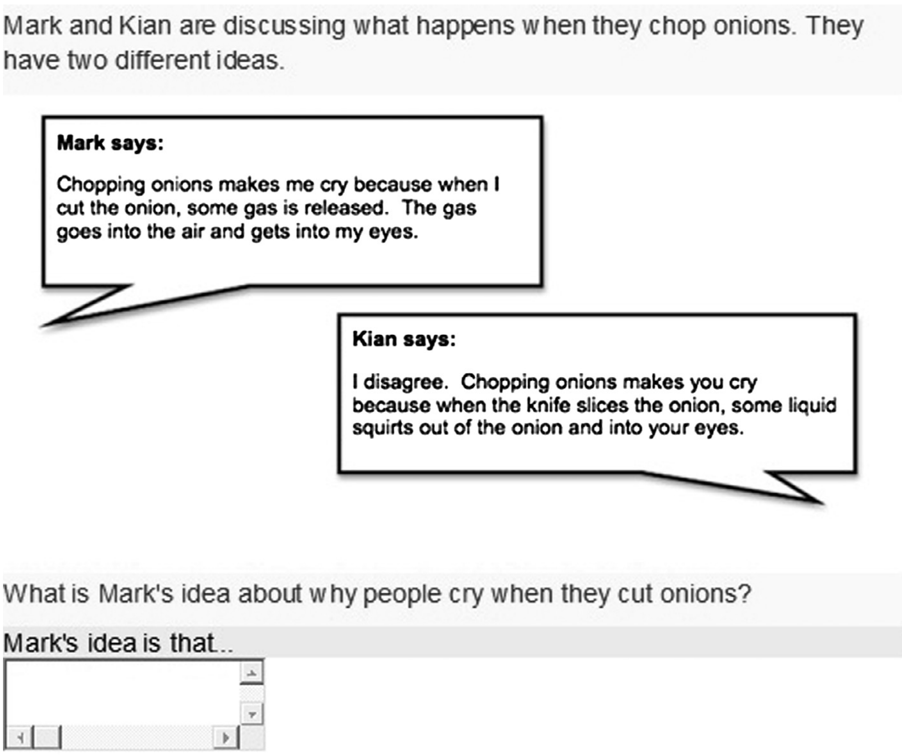

Fig. 1. An argumentation item from the *Onions* complex task.

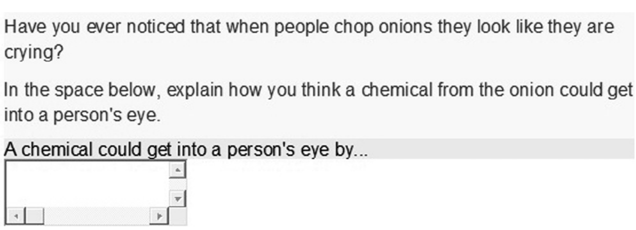

Fig. 2. An embedded content item from the *Onions* complex task.

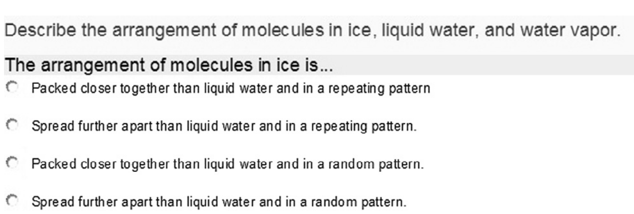

Fig. 3. A content item from the *Onions* complex task.

scientific context (e.g. what happens when you insert gas particles into a sealed container), and (3) content science items assessing knowledge of other concepts in the same science domain but not so closely associated with the context (e.g. compare the movement of liquid water molecules with the movement of ice molecules). In this paper, these three item types are referred to as 'argumentation', 'embedded content', and 'content.' Examples are provided in Figs. 1–3.

The 'complex tasks' are each set within a *context*—that is all the items within a complex task share a common setting. These contexts are what happens when someone (a) chops onions, (b) inserts gas particles into a container, and (c) mixes sugar into a glass of water. These contexts will be referred to as 'Onions', 'Gases', and 'Sugar' for easier reference. The embedded content and argumentation items were presented in these contexts, while the content items were related (i.e., they were more generalized but about the same concepts). Note that while the context of the content items are more generalized, they were still designated into a context by the test developers.

The remaining three item features explored in this paper are often tested in psychometric studies to examine whether they have an unintended effect on the item difficulties. For instance, the *format* refers to whether an item is open-ended or forced-choice (e.g., multiple-choice). Previous studies have suggested that multiple-choice items are easier for students than open-ended ones [10,11]. Because the assessment includes a combination of both, this feature is investigated to see if this finding holds true for the LPS data.

The *graphics* feature includes three categories: schematic representations, pictorial representations, and no graphics. Schematic representations are defined as abstract pictures whose "schematic meaning is provided by the symbolic/visual representation in the item" [13]. An example would include an image of the movement of gas particles. This contrasts with pictorial representations, which are concrete images that simply illustrate the details of objects described in an item.

Lastly, whether an item contains *academic words* is explored. Academic vocabulary words are those that are not among the 2000 most common words and occur most often in academic texts [4]. Unlike technical vocabulary—which are the specialized words specific to a discipline, academic vocabulary words are more generalized and span across many content areas [19]. This distinction is important for many studies investigating the language effects of content assessments because while technical vocabulary is deemed to be construct-relevant, academic vocabulary is often seen as construct-independent and, subsequently, may interfere with the interpretations of student scores on assessments [2,9,23]. Coxhead's [4] academic word list (AWL) is used here to identify academic words on the assessment[2]. Note that the word "evidence" is on the AWL, but will not be counted as an academic word in this paper because "evidence" is central to the argumentation construct. Thus, "evidence" is deemed to be construct-relevant, whereas other words on the AWL may be considered construct-independent.

[2] Cobb's website *Web Vocabprofile Classic* at http://www.lextutor.ca/vp/eng/ [3] automatically sorts texts and provides counts for four types of words: the 1000 most frequent words, 1001–2000 most frequent words, words on Coxhead's [4] Academic Word List, and off-list words.