Using Automatic Item Generation to Create Content for Computerized Formative Assessment

by

Xinxin Zhang

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Measurement, Evaluation and Cognition

Department of Educational Psychology

University of Alberta

Abstract

Computerized formative assessments (CFA) are becoming popular in post-secondary education because they can offer immediate feedback and on-demand testing. This has created a pressing need for large numbers of content-specified items together with their feedback so that the items may be available continuously for formative assessment. To support CFA in post-secondary education, I present a modified generation framework to address the operational difficulties that arise when applying existing frameworks. I also evaluate the application of the modified framework for item generation within the context of higher education. The proposed framework incorporates a human-machine interactive approach and employs a tree structure for cognitive modeling as well as introduces a new mechanism to assemble elements and a validation tool. The new framework was applied within the university teacher education field. Specifically, two instructors of EDPY 303 (Educational Assessment) used HIM_AIG, which is a software program developed based on the modified framework, to implement generation tasks for quizzes used in the course. Six experts in higher education evaluated the generated content using the statements that have achieved the greatest consensus among published item-writing and feedback-writing guidelines/recommendations. One hundred and thirty-four students took the practice quizzes, which contained the generated items. The results suggest that this modified approach is feasible; the evaluative quality of the generated items and feedback is comparable with that of the parent ones; the psychometric quality of the generated items satisfies the standard. Implications of the study and limitations of this research are also presented.

Preface

This thesis is an original work by Xinxin Zhang. The research project, of which this thesis is a part, received research ethics approval from the University of Alberta Research Ethics Board, Project Name “Using Automatic Item Generation to Create Content for Computerized Formative Assessment”, No.00084708, Oct 9, 2018.

Acknowledgment

First, I would like to express my sincere appreciation to my supervisor Dr. Mark Gierl for his continuous support of my Ph.D. study and related research, for his patience, motivation, generosity, and immense knowledge. I have been extremely blessed to have such an amazing supervisor.

In addition, I would like to thank my examining committee members Dr. Okan Bulut, Dr. Ying Cui, Dr. Hollis Lai, and Dr. Geoff Bostick all of the University of Alberta and Dr. Brandon LeBeau, University of Iowa for sharing their insightful comments and participating in my defense.

I was blessed to have so many friends in Edmonton. To CRAMERs, thank you for the continued support. To IFG friends, thank you for all the fun and tears we have had in the last years. To my church family, thank you for the prayers and blessings.

I would also like to express abiding love and gratitude to my parents, Ping and the late Yanpeng Zhang. When I was sad, the love of my husband Caixiang Fan, carried me through the dark moments. Thank you, Caixiang.

My deepest gratitude to my Lord Jesus Christ. Thank you for being so patient, faithful and merciful.

Table of contents

List of Figures

List of Tables

Chapter 1: Introduction

While educational assessment has typically followed a summative principle, it is now also embracing a formative principle (Miller, 2009). The traditional summative principle indicates that the purpose of assessment is to judge the students' performance based on marks, percentages, grades, or classifications (Cox, Imrie, & Miller, 2014). Summative assessment, which usually takes place at the end of a sequence of study, such as a unit, semi-term, term, or year, may take such forms as end-of-unit quizzes, midterm examinations, and certification tests. Unlike the summative principle, which focuses on making a judgment regarding students' performance at the end of a sequence of study, the formative principle focuses on providing students feedback – a specific rationale explaining why an answer is correct or incorrect – in order to stimulate learning throughout a sequence of study.

This shift from summative to formative assessment has been advocated by many educational researchers and practitioners (e.g., Black & Wiliam, 1998; VanEvera, 2003; Ruiz-Primo & Furtak, 2007; Hwang & Chang, 2011; Kingston & Nash, 2011) because of the positive effects of formative assessment on students' learning. For instance, Herman, Osmundson, Ayala, Schneider, and Timms (2006) claimed that "formative assessment should be considered prime among available interventions for improving student learning" (pp. 2–3). Dorn (2010) stated that "when used properly, formative assessment is one of the most powerful tools available to guide classroom decisions" (p.325). Specifically, these positive effects on students' learning are attributed to feedback given in a timely and detailed manner (Gierl & Lai, 2018; Van der Kleij, Feskens, & Eggen, 2015). However, while timely feedback, especially immediate feedback, is more efficient in improving learning (Corbett & Anderson, 2001; Mason & Bruning, 2001), and

is preferred (Miller, 2009) and valued (Van der Kleij, Eggen, Timmers, and Veldkamp, 2012) by students, it is challenging to implement. In particular, providing feedback in an immediate way on paper-and-pencil assessments is almost impossible because time is required to score the answers and then to present the corresponding feedback to students.

Computer-based assessment (CBA) that automatically conducts various assessment-related tasks through digital tools offers a possible solution since it can score the answers automatically and deliver feedback immediately. For example, when a student takes an online assessment with multiple-choice questions, as he/she completes each item, the answer is automatically compared with the key to generate a dichotomous score (0 means wrong, 1 means right). The corresponding feedback is immediately presented to the student, so that she/he can know how to correctly arrive at the right choice, and why other choices are implausible.

When this formative principle is combined with CBA, a new approach to assessment called Computerized Formative Assessment (CFA) emerges. CFA provides students with immediate feedback to support their learning while benefiting from the advances of computer technology (Miller, 2009). The advantages of CFA include immediate feedback and on-demand testing. These advantages may be especially valuable in post-secondary education (see Peat & Franklin, 2002; Burrow, Evdorides, Hallam, & Freer-Hewish. 2005) because they can address the purpose of and change in post-secondary education. One of the primary purposes of post-secondary education, as mentioned by Cox et al. (2014), is to "inspire and enable individuals to develop their capabilities" (as cited in Dearling, 1997, section 5.11). Timely feedback combined with the on-demand assessment can help achieve this goal, as empirical research has revealed that feedback with CBA items was at least moderately effective in supporting student learning (Miller, 2009). This approach also addresses a key change in post-secondary education: the fact

that there are more and more part-time students. The National Center for Education Statistics (2018) recently reported that the number of part-time students enrolled in US post-secondary institutions increased 15 percent between 2005 and 2015. During the same period, an increase of 39% in part-time enrolment and 10% in full-time enrolment occurred in Canadian post-secondary institutions (Canadian Association of University Teachers, 2018). For these students, it may be difficult to attend classes on campus. CFA provides these students with flexibility so that they can take the online assessment at any time and any place and, more importantly, they can receive valuable formative feedback throughout the course.

In order to support CFA in post-secondary education, large numbers of content-specific items, together with their rationales, are required so that the items may be available on a continuous basis for formative feedback. This demand imposes heavy burdens on subject-matter experts (SMEs) who are responsible for creating this content. To help students improve learning, SMEs are required to write large numbers of items that cover the instructional objectives in a given course so that students can evaluate themselves extensively and continuously. At the same time, the SMEs should write the corresponding rationales for each item so that students can receive feedback as they solve the item to reinforce the correct knowledge or identify their misconceptions. Currently, SMEs rely on a "one-at-a-time" process (Gierl & Lai, 2013): they write the individual items and develop the feedback specific to each item. This item development approach is time-consuming, particularly when large numbers of items and feedback are required.

Template-based AIG (LaDuca, Staples, Templeton, & Holzman, 1986; Bejar et al., 2002), creates items and feedback based on a template – a pattern that "highlight[s] the features in an assessment task that can be manipulated to produce new items" and feedback (Gierl & Lai, 2016a). This approach can ease these burdens because it provides an efficient way to generate

large numbers of content-specified items and rationales. However, for SMEs, there are some operational difficulties that arise when applying template-based AIG. These difficulties occur in phases such as content modeling, content generation, and content validation.

Purpose of Study

The purpose of my dissertation is to address these operational difficulties and explore the application of AIG to support CFA within the context of higher education. To achieve these goals, a comprehensive research study involving both theoretical and practical work will be conducted. Specifically, this work includes the following three parts: 1) identifying key problems by analyzing Gierl and Lai's AIG framework, 2) tailoring the AIG framework to CFA, and 3) applying this new framework for CFA in higher education.

Organization of the Dissertation

This dissertation will be organized into five chapters. The current chapter, Chapter 1, is an introduction to my research area as well as an overview and justification for this study. Chapter 2, a literature review, presents five key components of developing a CFA, providing additional details about content specification and content development (two components of the five). It also reviews AIG, which is an advanced approach in content development and identifies the strengths and limitations of the most comprehensive AIG framework. Chapter 3 describes my proposed a theoretical framework for addressing the limitations identified in chapter 2 and demonstrates the framework through a simplified example. Chapter 4 presents the application of theoretical framework and the results. Chapter 5 covers a discussion of the findings, a description of the limitations and future directions.

Chapter 2 Literature Review

Developing a CFA: Key Components

Developing a CFA requires a systematic approach. Building on the first edition of the "Handbook of Test Development" (Downing & Haladyna, 2006) and the 2014 "Standards for Educational and Psychological Testing" (AERA, APA & NCME, 2014), Lane, Raymond, Haladyna, and Downing (2016), in the second edition of the "Handbook of Test Development", outlined 12 components to provide a framework for developing and validating a test . These are as follows: 1) development of an overall plan, 2) definition of a domain and claims statements, 3) content specification, 4) item development, 5) test design and assembly, 6) test production, 7) test administration, 8) scoring, 9) definition of cut scores, 10) reporting test score, 11) test security, and 12) test documentation.

The authors claim all 12 of these "coordinated components are needed in the development of any test" (Lane et al., 2016, p. 3). However, some of the components are not suitable for CFA. For example, when the intended use of CFA is for practice, scoring might not be necessary. Thus, such components as cut scores, test score reports, and testing security are no longer necessary. Additionally, these 12 components do not include the development of feedback, which is the critical content for CFA. Therefore, following the Standards (AERA, APA & NCME, 2014), I have reorganized the 12 components (merging components 2 and 3, and components 5, 6 and 7; expanding component 4; and eliminating components 8,9,10, and,11) into a preliminary rubric of five necessary components that are directly relevant to the development of a CFA.

The first component is an *overall plan* that explicitly delineates the purpose and intended uses of a CFA and provides a systematic activities list. In this component, SMEs are required to address the fundamental questions such as: Who are the users? What are the intended uses of CFA? What is the construct or content domain to be measured? Based on the answers to these questions, SMEs plan activities such as specifying content and format, reviewing the specifications, developing items and feedback, and documenting the procedure.

The second component, *content specification*, includes "delineating the aspects (e.g., content, skills, processes and diagnosis features) of the content domain to be measured" (AERA, APA & NCME, 2014, p. 76) and defining the item formats and feedback types (Lane et al., 2016, p. 7). The delineated aspects, which extend the original statement of purpose and content domain developed in component one (AERA, APA & NCME, 2014), must be detailed and representative so that they can serve as a guide to develop items and feedback. The specification of item formats and feedback types must be aligned with the purpose and defined content (Wise & Plake, p. 21) to provide sufficient evidence of their content validity. For example, if the purpose of an assessment is to measure the ability of writing, item formats such as multiple choice and true-false may reduce the validity of the items as a measure of the intended purpose because they can not efficiently measure the writing ability.

The third component, *content development*, is most critical. Content development involves activities such as locating materials, creating items and feedback (to which we refer collectively as assessment content, or simply content, below), and obtaining evidence of validity to support the use of created content. These activities must align with the content specification described above. For example, if the content domain is medical education and the item format is

multiple choice, one of the suitable materials might be high-quality multiple-choice items which measure knowledge within a specific medical education context.

The fourth component is *test design and administration*. In this stage, SMEs address questions such as: What is the test form? Are there time limits? What are the accommodation procedures? What are the instructions and materials provided to users? How should the test-taking be monitored? Because the CFA coordinates both the items and the feedback through a computerized system, more complicated issues like presentation order, connectivity, and online administration should be considered.

The fifth component, *test documentation*, uses technical reports and other documentation to support the assessment's technical adequacy (Lane et al., 2016). The documented information includes but is not limited to content definition, item/feedback creation and review, and test administration. This documentation provides not only validity evidence to support the use of CFA (e.g., make claims about users' performance) but also a systematic record that can be easily updated for CFA with a new test content or purpose.

These five components are essential for guiding the development of a CFA, regardless of whether or not the results of formative assessments will be used for final grades. When the results of formative assessments are used by SMEs for determining students' final grades (Cox, Imrie, & Miller, 2014), additional components related to scoring, such as test score reports and test security, are taken into consideration. Among the five components, content specification and content development are particularly complex. The next sections will provide additional details about each of them, point out the challenges of the traditional approaches of developing content, and review AIG which is an alternative approach to address these challenges.

Content Specification

As noted above, both item formats and feedback types must be defined in this stage. The format of *items* includes constructed-response and selected-response types (AERA, APA & NCME, 2014; Albano & Rodriguez, 2018). The first type requires that the students construct a response. Many different constructed-response formats are available, including essay, computation, and fill-in-the-blank. The selected-response type requires the students to select one or more of the options, which are used by the majority of assessment today (Albano & Rodriguez, 2018). Within the selected-response types, formats can include multiple choice, true/ false, fill-in-the-blank multiple choice, and matching (Rodriguez, 2016).

In a CFA, the predominant format is selected-response, especially the multiple-choice type. An analysis of 35[1] empirical studies which were published between 1968 and 2011 and selected in a recent CFA study (Van der Kleij, Feskens, & Eggen, 2015) revealed that 74% of the studies (26) used multiple-choice items, as summarized in Table 1. The predominance of selected-response items in CFA corresponds to the general shift from using constructed-response to selected-response in the measurement world (Albano & Rodriguez, 2018). The possible reasons for this predominance are as follows: selected-response items are standardized, so it is more efficient and less expensive to score and provide feedback on them than on constructed-response items; CFA with selected-response items have an encouraging influence on student enactment (Wilson, Boyd, Chen, & Jamal, 2011) and are favored by students (Furnham, Batey, & Martin, 2011).

[1] Van der Kleij, Feskens, & Eggen (2015) selected 40 studies; only 35 of them provided information about item type.

Table 1

Summary of item type used in 35 computerized formative assessment studies

	Authors (year)	Item type		Authors (year)	Item type
1	Neri, Cucchiarini, and Strik (2008)	CR	19	Gordijn and Nijhof (2002)	SR(MC)
2	Morrison, Ross, Gopalakrishnan, and Casey (1995)	SR(MC)	20	Narciss and Huth (2006)	CR
3	Rosa and Leow (2004)	SR(MC)	21	Merril (1987)	SR(MC)
4	Epstein (1997)	SR(MC)	22	Moreno (2004)	SR(MC)
5	Xu (2009)	SR(Matching)	23	Murphy (2010)	SR(MC)
6	Cameron and Dwyer (2005)	SR(MC)	24	Mazingo (2006)	SR(MC)
7	Pridemore and Klein (1995)	SR(MC)	25	Nagata (1993)	CR
8	Ifenthaler (2010)	CR	26	Pridemore and Klein (1991)	SR(MC)
9	Clariana and Lee (2001)	CR, SR(MC)	27	Moreno and Valdez (2005)	SR(MC)
10	Lin (2006)	SR(MC)	28	Kopp, Stark, and Fischer (2008)	SR(MC)
11	Roos, Wise, and Plake (1997)	SR(MC)	29	Kramarski and Zeichner (2001)	CR
12	Papa, Aldrich, and Schumacker (1999)	SR(MC)	30	Collins, Carnine, and Gersten (1987)	CR, SR(MC)
13	Valdez (2009)	SR(MC)	31	Lee, Lim, and Grabowski (2010)	CR, SR(MC)
14	Hall, Adams, and Tardibuono (1968)	CR	32	Corbalan, Paas, and Cuypers (2010)	CR
15	Murphy (2007)	SR(MC)	33	Lipnevich and Smith (2009)	CR
16	Moreno (2007)	CR, SR(MC)	34	Kim and Phillips (1991)	SR(MC)
17	Munyofu (2008)	SR(MC)	35	Nagata and Swisher (1995)	CR
18	Sanz and Morgan Short (2004)	CR, SR(MC)			

CR= constructed response, SR= selected response, MC= multiple choice

Feedback is defined by Hattie and Timperley (2007) as "information provided by an agent (e.g., teacher, peer, book, parent, self, experience) regarding aspects of one's performance or understanding" (p. 81). In this dissertation, feedback is defined as information provided by the computerized system regarding learners' performances on a specific item intended to improve learning. There are various categorizations of feedback based on different standards, such as the functions of feedback (Black & Wiliam, 1998), its complexity (Narciss & Huth, 2004), the level of information it contains (Van der Kleij et al., 2015), and the time it is reported (Shute, 2008). In CFA, the choice of feedback type is affected by the item format. For example, if the item is a constructed-response form, the feedback probably is delayed because immediately scoring the constructed-response is challenging.

As mentioned above, selected-response items are most common for CFA. However, there are currently no specific recommendations for the feedback type of selected-response items. Based on a recent meta-analysis (Van der Kleij et al., 2015) on the effects of formative feedback in a computer-based environment on students' learning outcomes, a general recommendation for items in CFA suggests the items provide elaborated feedback (EF). EF is one of the three feedback types (Van der Kleij et al., 2015) classified based on the level of information. EF instructs the students beyond providing simple information about correctness (Van der Kleij et al., 2015; Shute, 2008). There are many forms of EF, such as worked examples, strategic hints, rationales, and study materials, among which rationales (see, Shute, 2008; Gierl & Lai, 2018), addressing the correct answer and/or explain why the selected response is incorrect, are suitable feedback forms for selected-response items.

The focus of this dissertation, accordingly, is on selected-response items (multiple-choice items) with elaborated feedback (rationale form) because the selected-response items are

most common in CFA and the elaborated feedback type is recommended as the most effective feedback type in CFA.

Content Development

To create the assessment content, which as noted above includes test items and feedback, three steps are needed. First, SMEs must locate suitable materials such as exemplary items, textbooks, or/and lecture slides. Next, they must develop the items and feedback. Finally, they collect evidence for content validity through evaluating the quality of materials and the created content.

Locating materials. The SMEs first need to locate the materials. Both the quality of materials and the degree of alignment with the content specification inform content validity (Wise & Plake, 2016). The high-quality materials including exemplary items, textbooks, and lecture slides (see, Rodriguez, 2016; Albano & Rodriguez, 2018). Novel materials such as news and research reports, depending on the discipline, can be used to create content. These materials should also provide important alignment with the content specification. Overly specific or general content should be avoided (Albano & Rodriguez, 2018). For example, if the content specification is particle theory of matter, materials about physics might be too general, while the ones about atoms might be too specific.

Creating content. Once the materials are located, the SMEs start creating the items and feedback. For selected-response items, SMEs create the stem, the options, and/or auxiliary information. The stem contains the statement and/or the question. The forms of options vary in terms of different item types. For multiple-choice items, the options include one correct option and multiple incorrect options. For true/false items, the options include one correct option and one incorrect option, and the values of these options are either “true” or “false.” Auxiliary

information includes images, tables, graphs, diagrams, audio, and/or video (Gierl, Zhou, & Alves, 2008). In addition to item content, SMEs also create feedback. As stated above, elaborated feedback (EF) type is recommended for selected-response items. Among the various forms of EF, the rationales, addressing the correct answer and/or explain why the selected response is incorrect, are most suitable because the specific information about responses can direct and facilitate students' learning (Shute, 2008).

The typical item content development approach is a "one-at-a-time" process. In other words, SMEs write the individual items and develop the feedback specific to each item. Ideally, SMEs should write individual content following detailed instructions which are usually found in item-writing guidelines (see, Haladyna & Rodriguez, 2013). A summary and reorganization of outcomes from four published item-writing guidelines from the last decade (i.e., Haladyna & Rodriguez, 2013; Moreno, Martínez, & Muñiz, 2006, 2015; Albano & Rodriguez, 2018) that include recommendations for statement, question, and options when creating educational assessment is presented in Table 2.

Table 2

Summary of item writing guidelines for developing statement, question, and options

		Moreno et al. (2006)	Haladyna & Rodriguez (2013)	Moreno et al. (2015)	Albano & Rodriguez (2018)
Statement	1.Present the main idea/intended content clearly	√	√	√	√
	2. Ensure that the linguistic complexity is suitable for the intended population of test takers.		√		√
Question	1. Word the question positively		√		√
Options	1. Use plausible and discriminating options	√	√	√	√
	2. Vary the location of the correct response according to the number of options		√		√
	3. Keep options independent/ Avoid giving clues to the correct response	√	√	√	√
	4. Avoid using the options all of the above, none of the above, and I don't know	√	√	√	√
	5. Phrase options positively		√	√	√
	6. Avoid the use of humor		√		√

Across the four studies, the rules with the most consensus in Table 2 are “present the main idea/intended content clearly,”which is for writing statement, “use plausible and discriminating options,” “keep distractors independent/avoid giving clues to the correct response,” and “avoid using the options all of the above, none of the above, and I don’t know,” which are for writing options. Other important rules with less consensus included “ensure that the linguistic complexity is suitable for the intended population of test takers,” which is for writing statement, “word the question positively,” which is for writing questions, and “phrase options positively,” which is for writing options. For the most part, the guidelines are consistent across studies, proving typical instructions for SMEs.

Table 3 contains a summary of the outcome from one systematic review, which analyzed more than 100 articles related to formative feedback (see Shute, 2008), and two meta-analyses, which focus on the effects of feedback in computerized environments (see, Van der et al., 2011, 2015). These three studies include guidelines/recommendations for developing feedback that enhances learning. Four of the 18 instructions summarized by Shute (2008) are directly related to rationale form. These four rules are “be elaborated,” “be specific and clear,” “present in manageable units,” and “minimize use of extensive error analyses and diagnosis.” Each of the rules is supported by other studies (e.g., Mason & Bruning, 2001; Narciss & Huth, 2004; Moreno, 2004; Williams, 1997; Kulhavy, White, Topp, Chan, & Adams,1985; Hoska, 1993). Furthermore, the systematic review (Van der et al., 2011) and meta-analysis (Van der et al., 2015) confirms the first rule “be elaborated.” Hence, SMEs can consider these guidelines for writing feedback.

Table 3

A Summary of feedback writing instructions

		Shute (2008)	Van der et al. (2011)	Van der et al. (2015)
rationales	1.Be elaborated (what, how, why)	√	√	√
	2. Be specific and clear	√		
	3. Present in manageable units	√		
	4. Minimize use of extensive error analyses and diagnosis.	√		

Validating the content. The third step is to validate the created items and feedback. Once the content is developed, validation is conducted by verifying the extent of compliance with the guidelines (Moreno, 2015). In addition , external reviewers (e.g., other SMEs) evaluate content for quality and clarity as well as accessibility (Lane et al., 2015; AERA, APA & NCME, 2014), thus avoiding introducing construct-irrelevant variance to the assessment (Kane, 2016; Lane et al., 2015). When developing content for CFA, content review rather than item tryouts (collecting psychometric properties) is often used because distinguishing students based on the psychometric properties of items (AERA, APA & NCME, 2014) is not the focus of CFA. The typical content review is used to determine domain clarity, evaluate the items on a test in relation to their relevance and representativeness, ensure the feedback is specific, clear and straightforward, and make sure no errors have occurred in the presentation of the items and rationales. Depending on the outcomes of these reviews, some content may be edited and reviewed again. Much like creating the content, the validation method is a 'one-at-a-time' process, such that SMEs need to scrutinize the content of each item and each piece of feedback.

Automatic Item Generation (AIG)

The typical content development approach described above is time-consuming and costly, particularly when large numbers of items and feedback are required to support the CFA. Automatic Item Generation (AIG) provides an alternative method. In this dissertation, AIG is defined as the task of automatically generating large numbers of content (e.g., statement, question, options, and feedback) using computer technology. AIG methods come in two forms: non-template-based and template-based. Non-template-based methods are closely linked with natural language processing, as they rely on artificial intelligence to generate content, while template-based methods are linked with content operation.

Non-template-based AIG. Relying on natural language processing techniques, non-template-based AIG can directly generate statement, question, options, and feedback from inputs such as text (in particular, declarative sentences) and knowledge bases. Existing non-template-based generation approaches can be classified into three categories: syntax–based, semantics-based, and sequence-based (Baghaee, 2017; Yao & Zhang, 2010). The syntax and semantics methods are conventional approaches to dealing with the content generation problem using linguistic rules, while the sequence-based approach is a recent and data-driven approach that uses the sequencing words or letters in existing content to predict and generate new content (Baghaee, 2017). Prior to examining each approach in detail, it will be useful to define the key terms.

As a part of grammar, *syntax* is a description of how words are combined in sentences (Jacquemin & Tzoukermann, 1999). The typical method to represent the syntactic structure of a grammatical sentence is with syntax tree or parse tree (Indurkhya & Damerau, 2010). *Semantics* focuses on the meaning of a sentence which is usually expressed through a logical and syntactical combination of words. Different syntactical combinations convey a different meaning (Briscoe, 2011). For example, the sentence "Sally gave (her) student bonus" and "Sally gave (her student) bonus" conveys different meanings. The first sentence emphasizes the bonus is a student bonus, while the second sentence emphasizes that the student is Sally's student. Although these two sentences look the same, due to the different orders of syntactic combining, they express different meanings. *Sequence*, in non-template-based frameworks, usually indicates a set of words or letters that follow each other.

In order to select appropriate approaches to generate content of computerized formative assessment, non-templated-based AIG literature has been grouped into four categories (statement, question, options, and feedback) based on the kind of content that is required. As Table 4

indicates, statements have been generated in a sequence-based approach; questions in syntax, semantics, and sequence-based approaches; options in a semantics-based approach; and feedback in a semantics-based approach. In the following sections, the representative research within the category will be reviewed and evaluated.

Table 4

Explored approaches in generating statement, question, option, and feedback

	Syntax–based	Semantics-based	Sequence-based
Statement			√
Question	√	√	√
Options		√	
Feedback		√	

Generating Statement. Davier (2018) is the only study that has adopted a sequence-based approach to generating statements. In particular, he used a multilayer long short-term memory recurrent neural network model, which is a deep learning model, to automatically generate short personality items (statements). Applying the model, which was trained from 3320 personality statements, he generated new items. He noted that 24 generated statements functioned well. However, he failed to provide a baseline. Thus, readers have no information on how many items were generated in order to produce 24 items. Furthermore, he claimed that due to the small size of his corpus (a corpus being defined as a collection of texts used for training the model), most of the generated statements are spelled correctly (Davier, 2018). However, this statement may not be accurate because a corpus must be large enough so that the trained model is accurate and generalizable. In the case of our study, if we were to adopt the sequence-based approach, we

would be required to provide a large corpus for training the model – a condition that could not be satisfied, as we do not have much content. Thus, in our case, the sequence-based approach may not be the right choice for generating statements.

Generating Question. Generating questions has gained the most attention from the non-template-based researchers, most of whom are from the discipline of linguistic computing. They have explored all three approaches: syntax, semantics, and sequence-based. They have even given this task a name: Automatic Question Generation (AQG) (Rus, Graesser, & Cai, 2008). In the following sections, a review of question generation is presented.

The syntax-based approach works at the syntax level, usually representing languages and defining transformation with syntax (Yao, Bouma, Zhang, Piwek, & Boyer, 2012). The standard syntax-related techniques include but are not limited to part of speech tagging, defining syntactic categories (verbs, noun phrases), and constructing a syntax tree. In order to provide a rationale for not choosing this approach, a misconception is first noted and then a review of two classical and advanced studies is presented.

Gates (2008) combined several existing pieces of software, such as IdentiFinder (Bikel, Schwartz, & Weischedel, 1999), ASSERT (Pradhan et al. 2005), and Stanford NL parser (Klein & Manning 2003), to generate questions from sentences in a reading passage. The workflow of his framework involves first forming a syntax tree from the input sentence, then transforming the syntax tree into a Wh-question tree based on human-defined tree transformation rules, and finally generating questions from the Wh-question tree. The acceptable rate of the generated questions was 81%, based on an independent rater who rated each question in terms of the grammaticality and the practical usefulness of the questions. Similarly, to generate questions from reading passages, Heilman and Smith (2009) developed a three-stage syntax-based approach including

these steps: transforming source sentences, generating questions, and ranking questions. Their implementation achieved 43.3% acceptability for the top 10 ranked questions based on 15 judges who indicated whether the generated questions exhibited any of deficiencies, such as: being "ungrammatical", not making sense, being vague, having an "obvious" and a "missing" answer, using the "wrong wh word", or exhibiting a formatting error (Heilman & Smith, 2009, p 13).

This large difference in acceptability rates is not only due to the different frameworks used, but also to the variation in corpora and the methods of evaluation. Gate (2008) used a domain-specific corpus for training and testing and evaluated results through a single human judge with only two grammatical standards, while Heilman and Smith (2009) used four domain-general corpora and evaluated results through fifteen judges with seven grammatical standards. Although the results of Heilman and Smith (2009) are not compelling, they are most accurate for evaluation the quality of the generated items. Unfortunately, in some published studies, the researchers evaluated the generated questions according to the biased standards as mentioned above. So, unsurprisingly, the misconception that natural language processing is very mature in generating content is put forward in the literature.

Based on Heilman and Smith (2009), Heilman (2011) in his dissertation presented a three-stage process for factual question generation, which is deemed a state-of-art syntax-based framework (Danon & Last, 2017). In the question creation (second stage) of a three-stage framework, he conducted a series of operations at the syntax level, such as marking not suitable phrases, selecting suitable ones, decomposing the main verb, inverting the subject and auxiliary verb, and inserting one question phrase at the beginning of the main clause (Heilman & Smith, 2009). The overall acceptability rate of the top-ranked questions was around 40-50%, varying with different sources of input.

Building on Heilman's system (2011), Danon and Last (2017) added two more steps: paraphrasing the verb and identifying the hypernym. This new framework modified 148 rejected questions out of 217 top questions that were generated by Heilman's system (2011). Sixty-two questions among the 148 were accepted by two rankers (achieving an acceptance rate of 42%). Although Dannon and Last's framework addressed the limitation of Heilman's, its rate of acceptance was still low. This low acceptance was mainly due to incorrect syntactic structures and non-relevance to the main topics in the text (Danon & Last, 2017). Unless these two problems can be solved, it is unpractical to apply a syntax-based approach to generating questions because the quality of generated items is too low.

The semantics-based approach generates questions at the semantic level. Semantics related techniques include finding synonyms (Gütl, Lankmayr, Weinhofer, & Höfler, 2011; Singh Bhatia, Kirti, & Saha, 2013), word sense disambiguation (Brown, Frishkoff, & Eskenazi, 2005; Susanti, Iida, & Tokunaga, 2015), and translating from one natural language to another. Like the syntax-based approaches, semantic-based approaches rely on the tools, particularly parser, for analyzing structure. However, semantic parsers are typically more error-prone than syntactic ones (Heilman, 2011). As a result, the performance of the semantic-based approach is usually worse than syntax-based approach, though the approach seems more straightforward and high-level.

A small number of studies have explored the semantic-based approach for question generation and pointed out the limitations of currently available tools to present semantic structure. For example, Mannem, Prasad, and Joshi (2010) generated factual questions from paragraphs through a semantic role label which identifies semantically appropriate parts of a sentence. They claimed that the accuracy rate of semantic role labeler used in their work was less

than 85%. In another study, Yao and Zhang (2010) adopted a meta-level language for describing semantic structures to generate items from paragraphs. Although meta-level language is an improvement over other methods, they pointed out that this meta-level language was still unreliable, causing the failure of generation. Two years later, Yao, Bouma, Zhang, Piwek, and Boyer (2012) presented an updated semantics-based system that generated questions from only a single declarative sentence. Based on the criteria (relevance, question type, syntactic correctness and fluency, ambiguity, and variety) of the Question Generation Shared Task and Evaluation Challenge (QGSTEC, 2010; Rus et al., 2010;), their system achieved a total of 8.6 (5 is the best and 16 is the worst). They claimed that the implementation of semantic approach is limited to tools that are currently available. Unless the tools such as the parser, and the preprocessors improve, the application of the semantic approach to generate questions is limited.

The Sequence-based approach is often viewed as the non-template-based method with the most potential for generating questions. Recently, researchers have shifted their attention from the traditional approaches to this sequence-based approach, particularly to Recurrent Neural Network (RNN), which is a deep learning model. Compared with syntax and semantics-based approaches, sequence-based generation relies less on linguistic structures, and more on mapping content to vectors of numbers. In this section, I review multiple sequence-based studies that used RNN to show the logic and limitations of this approach for solving our problem.

Serban et al. (2016) formed a corpus from Simple Questions that is the largest dataset of question-answer pairs and trained it with RNN. They used both automatic evaluation metrics, which measure the quality of generated content with machine, and human evaluators to evaluate the performance of their framework. All evaluation metrics (e.g., BLEU, METEOR, and Emb. Greedy) indicated that their models outperformed the baseline model. The pairwise preference

experiment with human evaluators, who were asked to select the most relevant question from a pair of questions (one human-created and one generated one), showed 93% of the generated questions were either better than or at least as good as the human-created ones. In another study, Baghaee (2017) trained a long short-term memory RNN model on 912255 questions retrieved from Amazon question/answer dataset to generate new questions. Regarding syntactic correctness and relevance, his system separately achieved 4.52 out of 5 and 3.37 out of 5 from human judges (5 means perfect). Using reinforcement learning, Yuan et al. (2017) trained an RNN model for question generation on 87599 question-answer pairs. To evaluate the quality of the generated questions, they compared the performance of their model with the baseline model using five evaluation metrics (NLL, BLUE, FI, QA, and PPL), claiming their model performs better as all scores of the metrics improved.

These results suggested a potential value of sequence-based approaches for generating questions. However, in the case of our study, applying sequence-based approaches to question generation will induce a paradox because the prerequisite for applying such approaches – the requirement of a large amount of existing questions (for training) – is the problem this study is seeking to solve. Thus, for the purposes of current study, sequence-based approaches are not adopted for generating questions.

Generating Options. The semantic-based approach is mainly adopted to generate options because the options are usually short words or phrases which have no syntax structure and lack a sequence. Thus, alternatives – syntax-based and sequence-based approaches are inappropriate. Mitkov and Ha (2003) used a semantic-based approach to automatically generate options. They identified nouns or nouns phrases with a high frequency from the input text as the key and then selected semantically close concept through Wordnet as distractors. The results show that 6 out

of the 65 generated distractors were poor, as they either attracted more students from the upper group than from the lower group or had no student to choose; while in control group, 12 out of 33 human-developed distractors were poor. Although this result suggests the potential of applying semantic-based approach for generating options, for our problem, applying it is inappropriate because not always the nouns with high frequencies can be the keys and not always the semantically close concept can be reasonable distractors.

Generating Feedback. It is rare to see researchers generate feedback with non-template-based approach. Fossati, Di Eugenio, Ohlsson, Brown, and Chen (2015) adopted a semantic-based approach to generate reactive procedural feedback, which is one of the five types of feedback they generated for a procedure assessment (these fives types include syntax, execution, final, reactive procedural, and proactive procedural type). This inference is based on their description of the workflow. To give students feedback, the system first records the students' past state, then compares it with the current one, and finally summarizes the differences between those. Summarization is very possible to be conducted through organizing information at the semantical level. This semantic-based approach is not adaptable to solve our problems because it was used to generate feedback for procedure assessment, which is not the focus of this dissertation.

To summarize, researchers mainly from linguistic computing area are actively exploring non-template-based approaches for automatically generating content. Although these approaches are computationally advanced, they are inappropriate to solve our problem – generating large numbers of content for CFA – for two main reasons: first, implementing syntax-based and sematic-based approaches is limited to current available tools such as those extracting the syntax and semantic structure. Therefore, the overall quality of generated content based on existing

syntax-based and semantic-based approaches are low; second, sequence-based approach requires a large corpus for training the model – a condition that could not be satisfied, as the challenge this study is seeking to solve is creating large numbers of content. In the next section, I will review another approach called template-based AIG.

Template-based AIG. Template-based AIG can be traced back to 1957 when Guttman suggested using common component of knowledge to create multiple test items (as cited in Lai, 2003, p.15), but systematic research did not become popular until the 1990s (Papanastasiou, 2003). Likewise, template-based research has been grouped into four categories: statement, question, options, and feedback. Within each content category, a review of the related research is presented.

Generating Statement. The template-based method for statement generation is intended to create statements with comprehensive content. Researchers adopted different templates for specific purposes. For example, to automatically generate statements of parade ground items, which described a route in the plane and asked a question about the direction of flight, Dennis (2002) used the template presented in Figure 1. This template lists controlling features whose variation affects item difficulty and non-controlling features whose variation do not affect item difficulty. Items are produced by the factorial combinations of a specified set of features.

Statement of a parade ground item

A yacht leaves its mooring, sails 10 km North-East, a further 15 km South-East, then another 4 km North East and finally 15km North-West.

Template

Controlling features: Use of cardinal VS semi-cardinal compass points

Noncontrolling features: Scenario

Figure 1. An example of Dennis (2011) approach to generate statement.

To generate statements of quantitative-comparison items, Bejar et al. (2002) applied templates with constraints which confine the possible values of the features. For example, the feature "area of a triangle" is constrained to be equal to the length of vertical leg multiply ½ of the length of the horizontal leg of the triangle. To measure higher-order cognition, Hornke (2002) provided a verbal analogies template to generate statements. These template-based methods made important contributions to specific questions. In 2013, Gierl and Lai published a generalized framework for generating multiple-choice items. (Although this framework is for multiple-choice items, one of its core functions is to generate statements. So it is reviewed here.) This framework has three stages. The first stage is to produce a cognitive model which is a structured representation of the cognitive- and content-specific information required to solve questions on tests. The second stage is to produce item model which is a template to embed the cognitive model. The third stage is to generate items using a generator.

The template-based approach for statement generation overcomes the limitations of non-template-based approaches, such as large numbers of existing statements are required for training model as a prerequisite; the theoretical work is lacking; the generated statements are not accurate. Compared with non-template-based approach, template-based approach has no prerequisites; has

abundant theoretical work that have been applied for different purposes, such as recruitment test (e.g. aviation and space test to select pilots/ astronauts) (Goeters & Lorenz, 2002), admission test (e.g. GRE), language test (e.g. word fluency test) (Arendasy, Sommer, & Mayr, 2012), license exam (e.g., license of architects) (Gierl et al., 2008;) and classroom assessment (Gierl, Bulut, & Zhang, 2018). In addition, the generated statements are more comprehensive and accurate (Zhang & Gierl, 2018). For these reasons, the templated-based approach for statement generation will be employed.

Generating Question. The template-based method for question generation usually relies on a pre-defined template derived from the grammatical structure of factual statements. The template-based approaches rely on the idea that a question template can capture a class of context-specific questions having a typical structure. Research has been conducted to explored how templated-based approaches can be applied to generate questions. Bormuth (1970) put forward a grammatical transformation method for creating questions: statement structure is first abstracted and then question structure is derived from the statement structure based on the grammatical rule. Chen, Aist, & Mostow (2009) applied handcrafted-templates such as "What would happen if <X>?" and "When would < X > happen?" to produce questions from narrative fiction, where < X > is the feature of the template (see also Mostow & Chen, 2009; Lindberg, Popowich, & Winne, 2013; Ali, Chali, & Hasan, 2010).

Compared with non-template-based approach for question generation, this template approach is robust (Yao and Zhang, 2010) because the pre-defined templates can ensure the high quality of generated questions. However, the cost is relatively high as human labor is needed for creating the template. In the current study, fixed questions will be used, so neither non-template-based or template-based are considered.

Generating Options. Lai et al. (2016) described a mechanism for creating distractors where distractors share some of the all features with the key. These distractors are not automatically created through a template but are predefined by the SMEs. Other literature used templates to create options. For example, Gierl, Zhou, and Alves (2008) showed that options could be generated from the sub-templates such as "I1+I2+I3+I4", where I1, I2, I3, and I4 are integer features. In addition to mathematics, nonverbal reasoning is another area that templates are used to generate options. For example, Gierl, Ball, Vele, and Lai (2015) presented a n-layer model, which is a template with hierarchical structure in which elements on multiple levels can be varied, in order to generate diverse options for nonverbal reasoning items.

Compared with non-template-based approach for options generation, this template approach is more suitable for the structural options. In other words, if these options can be structured by either a mechanism or a template, template-based approach might be an excellent choice. Usually, the options required for computerized formative assessment are structural because these options can be arrived at either through a formula, algorithm, and/or a logical structure, or by combining the features of a task. Hence, in this study template-based approach will be used to generate options.

Generating Feedback. The first trial of applying templated-based method for feedback took place in 2016, when Gierl and Lai embedded rationales generation into their general AIG framework. This method has two stages. The first stage is to produce a rational model which contains three types of rationales and elements. The three types of rationales are key feature list rationale, key feature set rationale, and key feature set with distractor rationale, separately focusing on presenting all of the key features, emphasizing specific combinations of key features,

and explaining why the distractors yield the incorrect solution. The second stage is to generate feedback based on the rational model.

Compared with the non-template-based approach for feedback generation which is designed specifically for procedure assessment, the template-based approach suits the scope of the current study, which is involves general assessment. Furthermore, the feedback generated based on this approach is elaborated and specific, satisfying the requirement of current study. Additionally, as noted above, this feedback generation approach can be imbedded into existing template-based approaches which will be used for generating the statement and options. For these reasons, a template-based approach will be adopted in current study to generate feedback. Among all template-based approaches, moreover, Gierl and Lai's framework is chosen for current study, because it is the most comprehensive one covering the generation of statement, questions, options and feedback. Next, I will review Gierl and Lai's framework and point out some limitations in practice. The purpose of this dissertation is to address these limitations.

Gierl and Lai's AIG framework

Based on Gierl and Lai's past work (e.g., Gierl & Lai, 2013, Gierl & Lai, 2016b, Gierl & Lai, 2016c), their template-based framework can be summarized in three stages: content modeling, content generation, and content validation. These three stages match the standard three-step procedure of content development: locating materials, creating content, and validating content, as discussed in Chapter 2.

Content Modeling. Content modeling is a process for creating an item model and feedback model. To develop an item model which contains the components (such as statement, question, options, etc.) in an assessment task, a cognitive model must be derived from a representative item called parent item. Gierl and Lai (2017) described two types of cognitive

models – logical structures and key features. While logical structures cognitive model is applied "when a specific concept or a set of closely related concepts is operationalized as part of the generative process." (Gierl & Lai, 2017, p.131), the later one "is used when the features of a task are systematically combined to produce meaningful outcomes across the permissible feature set." (Gierl & Lai, 2017, p.135). The focus of this study is on the key features cognitive model as the existing methods of developing distractors and feedback are derived from it. The cognitive model (Figure 2) consists of three panels: problem and scenario, source of information, and elements and constraints.

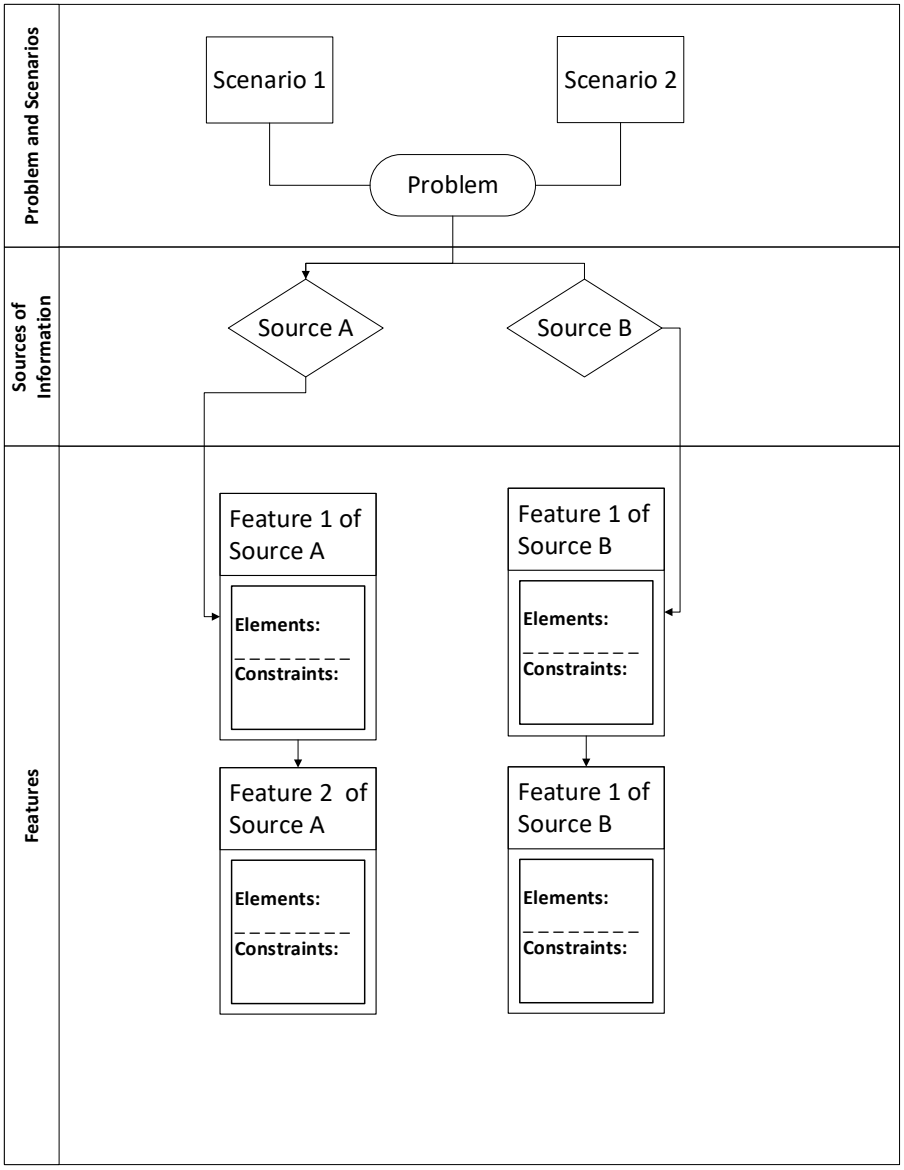


Figure 2. An illustration of a cognitive model (Gierl and Lai, 2013).

This cognitive model is then used to guide the development of an item model (Table 5) which has three components: stem, elements, and options. The top box (stem-box) presents a stem with variables. The middle box (element box) shows the corresponded content for these variables. The bottom box (option box) lists all the options including keys and distractors which can be created through a method called systematic distracter generation (Lai, 2016) that specifies the key features related to errors and misconceptions to generate plausible but incorrect options.

The generated distractors share some but not all of the key features required to identify the keys. The shared features with the key make the distractors plausible, and the unique features make the distractors incorrect.

Table 5

An item model

Stem	… [Feature 1 of source A] … [Feature 2 of source A] … [Feature 1 of source B] … [Feature 2 of source B] …
Elements	[Feature 1 of source A]: values set 1 [Feature 2 of source A]: values set 2 [Feature 1 of source B]: values set 3 [Feature 2 of source B]: values set 4
Options	key1; key2; distractor 1; distractor 2; distractor 3

A feedback model (Table 6) contains the rationales and elements. The upper box (rationale-box) presents a rationale which provides an explanation for both the key and distractors of an item. The bottom box (elements box) shows the corresponding values for these variables. Some of these values are cited directly from the item model, like the values of [Feature 1 of source A], [Feature 2 of source A], [Feature 1 of source B], and [Feature 2 of source B]. Some of them are newly coded, such as [Distractor Rationale] which is coded to present the common elements among the distractors as the explanation for why the distractors would be erroneous responses.

Table 6

A feedback model

Rationale	[Key] is the correct option as … [Feature 1 of source A] … [Feature 2 of source A] … [Feature 1 of source B] … [Feature 2 of source B]. The distractors are not plausible because [Distracter Rationale]
Elements	[Feature 1 of source A]: values set 1 [Feature 2 of source A]: values set 2 [Feature 1 of source B]: values set 3 [Feature 2 of source B]: values set 4 [Distractor rationale]: values set 5 (newly coded)

Content Generation. Content generation is the process of creating items and feedback by placing the variable content (elements) into the models with computer systems. In this process, all possible combinations of the variable content are assembled subject to the constraints which ensure the generated items and feedback are sensible and useful. For items, the constraints have been articulated in the cognitive model, while for feedback, the constraints must be identified through extra work, such as organizing and analyzing the features lead to both distractors and keys, since there is no representative structure (e.g., a cognitive model for feedback) to refer. Gierl, Lai, and Matovinovic (in press) introduces how IGOR computer system (Gierl et al., 2008) functions to generate items: SMEs enter the components (stem, variables, options) of an item model in separate panels and then specify the size and save location of the generated items bank. Specifically, with a new version IGOR, constraints between variables are automatically coded as "a binary, element-by-element, matrix" (Gierl, Lai, & Matovinovic, in press, p.12) to procure new items, as SMEs set the constraints by checking boxes.

Content Validation. Model review is an alternative content validation method for generated items (Gierl & Lai, 2016a). The logic behind this method is if the instructions for generation in the models are sound, then the outcome from the generative process (i.e., the individual items) should also be sound. Two rounds of reviews are needed for content validation. The first round focuses on evaluating the quality of the AIG models including cognitive and item model. During the first review, the SME who developed the AIG models is responsible for evaluating the content and the logic. The second-round review is conducted by other SMEs who did not develop the AIG models. These SMEs will evaluate the content and the logic as well as provide feedback for model improvement using a standardized rating scale.

Limitations. Gierl and Lai's comprehensive framework provides a potential solution to our problem. However, there are still some limitations, which are summarized here according to the stages of the content development process.

In the *content modeling* stage, the most significant limitation is related to effort and time. SMEs are required to make much effort to develop the cognitive, item, and feedback models. In the case of the cognitive model, content experts need to spend a significant amount of time in training and practice, so that they are able to express their knowledge in the proposed format (e.g., expressing the condition for each feature in the form of constraints). As Hoffman and Lintern (2006) point out, "No matter how much detail is provided about the conduct of a knowledge elicitation procedure, there is no substitute for practice." Likewise, in developing an item model, SMEs need to extract the information from the cognitive model and then organize them appropriately – a process that takes time. The same is true of developing a feedback model. SMEs must take a series of extra steps, such as collecting

specific information for all of the keys and distractors, identifying the shared features of distractors, coding corresponding distractor rationale, and setting constraints.

Other limitations are specifically related to the feedback model. The current approach to developing such a model assumes that the distractors have common elements. Hence these common elements, which produce an incorrect option, are presented to the student to explain why the distractors are implausible. Suppose, for example, distractor 1, distractor 2 and distractor 3 are all associated with feature 3 of source B (the key is only associated with feature 1 and 2). They are thus presented as distractors of one item with the rationale "distractors are implausible because none of them should appear with feature 3." However, this assumption is not always practical because distractors may not share common elements. Furthermore, the current feedback model approach is not individualized. In other words, no matter which option students choose, they always receive the same rationale. To better help students learn, an approach for generating individualized feedback (a rationale for each option) is needed.

In the *content generation* stage, as the content experts are required to program the model, time also presents a challenge. To generate items, for example, SMEs need to first identify constraints pre-defined in the cognitive model and then set constraints in a way the computer system accepts (e.g., entering constraints as a boolean logic form or checking boxes). To generate feedback, they need to analyze the features of both distractors and keys, identify the constraints, and set these constraints in a generator. In addition to setting the constraints, SMEs manually input the stem, elements, and options into the generator, which is tedious.

In the *content validation* stage, the current method relies on the assumption that SMEs correctly program the constraint logic when they input the item model into the generator. However, in practice, this is not always true. Thus, the current validation method cannot detect the errors SMEs make when they program the model. Furthermore, a method for validating distractors and their rationales is lacking.

To conclude, the literature review presented in this chapter reveals that while template-based approaches are most appropriate to serve the purpose of the current study, the most comprehensive framework is template-based one as described by Gierl and Lai. Despite the merits of this approach, template-based AIG still has limitations. The next chapter proposes a new framework which overcomes these limitations by employing tree structures as a mechanism to create content in a more efficient way.

Chapter 3 Methodology

In this chapter, a proposed framework for item generation is presented, and one example is offered to demonstrate how items are produced based on this framework. This modified methodology is intended to address the limitations of Gierl and Lai's framework. It comprises three stages – *content modeling, content generation, and content validation* – and can be used to achieve key goals related to each of these stages. First, it can simultaneously create cognitive, item, and feedback models without requiring extra inductive work from the subject matter experts (SMEs). Second, it can automatically generate items and rationales for individual options from the created models without the need for programming. Third, it can provide a comprehensive structure for content validation.

Framework

Content modeling. Content modeling is a process of creating models – such as cognitive models or maps of knowledge required to solve problems, item models or templates for generating items, and feedback model or templates for producing feedback – to guide the generation of content. Gierl and Lai's framework adopts a linear approach to this process. In other words, SMEs must develop a cognitive model first, and then develop an item model based on the cognitive model. They then transfer the item model into IGOR software to generate items. Gierl and Lai's framework also adopts a segmented approach, meaning SMEs create an item model and feedback model independently. This may be inefficient given that, when SMEs create a feedback model, they usually need to manually create more variables and their corresponding values, which requires both time and effort. To improve the efficiency of using AIG, the modified framework will adopt a parallel and connected approach through

human-machine interaction. This approach can create a cognitive model, item model, and feedback model in one step.

In our framework, a particular data structure, referred to as a *tree structure*, is used to develop the cognitive model, given its capacity to capture both concrete content and abstract structure for generating items and feedback (see, Zhang & Gierl, 2016). More importantly, thanks to the special attributes of a tree structure, elements, which are connected via nodes and edges, can be automatically and logically searched and combined in a process called tree traversal. Thus, SMEs do not need to set constraints on how to combine them. Before we present the details of methodology, necessary information about a tree structure will be introduced.

A basic review of the tree structure. A tree is a data structure, which is an organization and storage format that collects data values and specifies the relationships among them. A tree normally consists of a root node (top node) and potentially many levels of additional nodes, which are connected by edges to form a hierarchical form. Figure 3 shows a tree, where the top node is a root that has three nodes directly connected to it; each of the three nodes has two connected nodes when moving away from the root. These six nodes at the third level are called *leaves*, as they have no nodes beneath. A *forest* is a set of $n \geq 0$ disjoint trees. Figure 4 presents a forest with two disjoint trees. We use these trees to depict the cognitive model and use tree traversal, which involves visiting each node in a tree data structure, to automatically combine elements. There are two ways to traverse a tree: via a depth-first search, which explores as far as possible along each branch before backtracking, or through a breadth-first search, which explores all of the neighbor nodes at the present level prior to moving on to the nodes at the next level.

For the tree on the top in Figure 4, the depth-first and breadth-first search give the same result – ABCD. In this way, the information stored in each node is presented in a logical way.

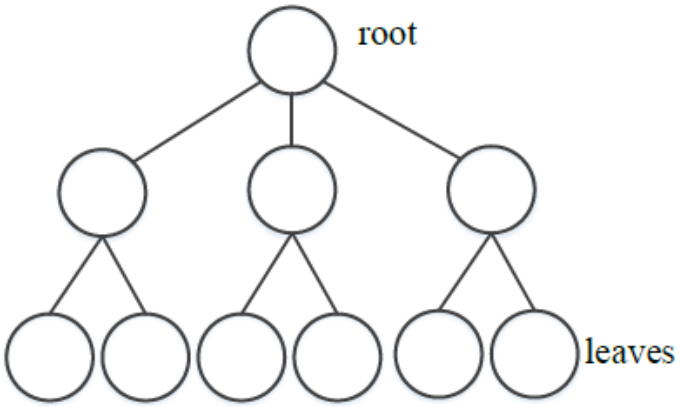


Figure 3. A tree structure.

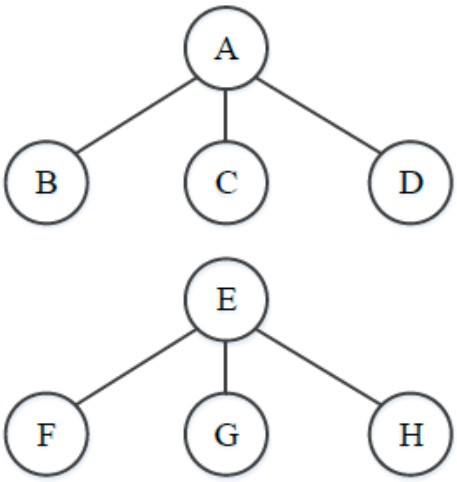


Figure 4. A forest with two disjoint trees.

These disjoint trees can be transferred into an integrated tree structure – a binary tree (Figure 5) – on which each node has fewer than two children and/or one parent. In this study, the binary tree will be used as an alternative structure to present the cognitive model.

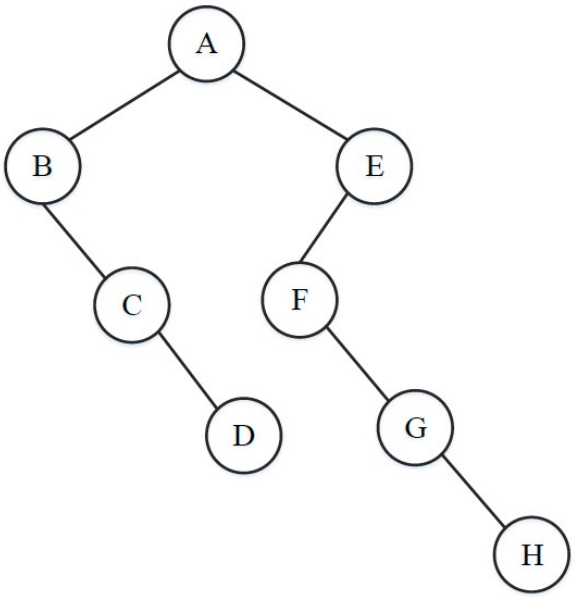


Figure 5. A binary tree

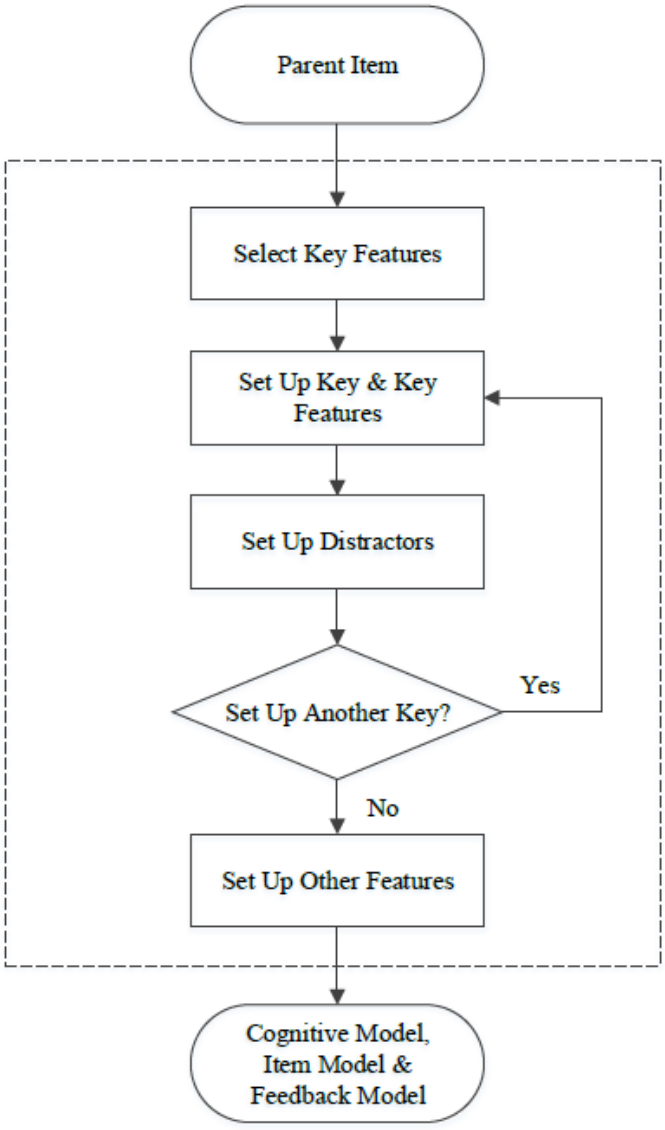


Figure 6. Workflow of creating content models.

Workflow involved in creating content models. Figure 6 shows the workflow involved in creating content models. It presents a human-machine interactive process in which a human (SME) continually interacts with digital tools (computer) through graphical user interface to collaboratively complete a task (creating content models). The starting point is to use a parent item or an exemplar that is well written and of high quality based on the judgement of SMEs.

The parent can be found by reviewing items from previous test administrations, by drawing on a bank of existing test items, or by creating the parent item directly. This item helps SMEs to determine the underlying structure of the model, thereby providing a point-of-reference for creating alternative items (Gierl & Lai, 2016).

The first step of the human-machine interactive process involves asking SMEs to identify the key features, or pieces of information, that are needed to solve a parent item. In this step, SMEs identify key features that can be systematically combined to produce meaningful outcomes by reasoning about how to solve the item. These features play a role in developing each type of content model. For the cognitive model, an initial tree structure (Figure 7) is generated, where the root is the problem, and the leaves are associated key features. For the item model, the key features of the parent item are automatically replaced with abstract variables forming the stem of an item model. We call these variables *key feature variables*. The feedback model, by default, is defined as: [key] is correct, because it matches every feature [all key features]; and for each distractor: [distractor] is incorrect, because it does not match the following feature(s): [certain key features].

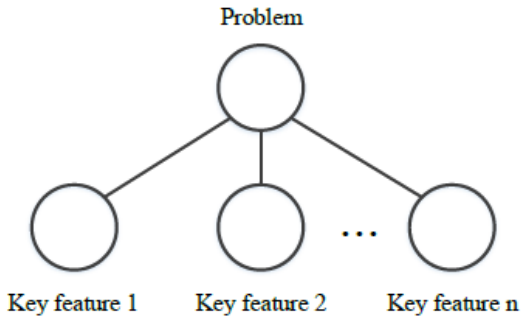


Figure 7. An initial tree.

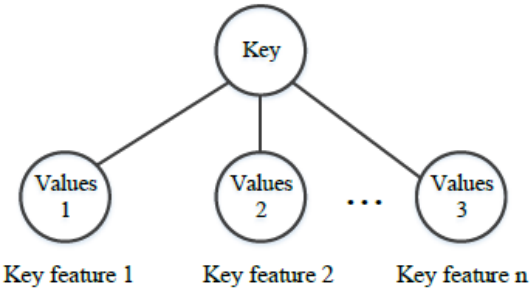


Figure 8. A tree with inserted values.

The second step in content modeling is to identify a scenario (key) and provide values corresponding with each of the key feature variables created in step 1. All combinations of these values should lead to this key. They will be used to produce new items and feedback. For the cognitive model, the key is inserted into the root of the tree produced in step 1 and the values of key features are inserted into corresponding leaves (Figure 8). For the item model, the values for key and key feature variables are presented.

The third step is to define distractors. Two question templates are designed to generate questions that help to guide SMEs in identifying plausible distractors; SMEs may use either of the templates. The first template is: "Can you identify distractors that do not match the following feature(s): [all values of a key feature] but match at least one of the feature(s): [the values of the rest key features]?". The values for replacing the variable [all values of a key feature] of this template are the values that SMEs had entered in step 2 for each of the key features, and the variable [the values of the rest of key features] includes all values for the rest of the key features. In Figure 8, for example, the questions may be "Can you identify distractors that do not match the following feature(s): [Values 1] but match at least one of the feature(s): [Values 2, Values 3…Values N]?" or "Can you identify distractors that do not match the following feature(s): [Values 2] but match at least one of the feature(s): [Values 1, Values 3…Values N]?" The second template is: "Can you identify distractors that do not match the following feature(s): [one value of a key feature] but match at least one of the feature(s): [the values of the rest key features]?" In Figure 8, for example, the questions can be "Can you identify distractors that do not match the following feature(s): [one element of Values 1] but match at least one of the feature(s): [Values 2, Values 3…Values N]?" In practice, template 1 is recommended for use when saving time is the priority; template 2 is recommended when more diverse distractors and

comprehensive feedback are required. More specific information can be found in the section "Demonstration with an Example".

In the third step, SMEs type the answer (distractors set) for the questions generated from either of the templates. In this way, the information about how each distractor relates to the key in terms of key features is known, which can be used for generating feedback. For the cognitive model, these answers are inserted into the tree as the children of the original leaves (key features) to show how they are related to the key features of key (Figure 9). Figure 9 is a formed tree when template1 is used. As this tree illustrates, distractor set 1 does not share all values of key feature 1 with the key; distractor set 2 does not share all values of key feature 2 with the key. Figure 10 shows a formed tree when template 2 is used. As it describes, distractor set 1.1 does not share the values 1.1 of key feature 1 with the key; distractor set 2.1 does not share the values 2.1 of key feature 2 with the key. For the item model, these answers are presented as distractors.

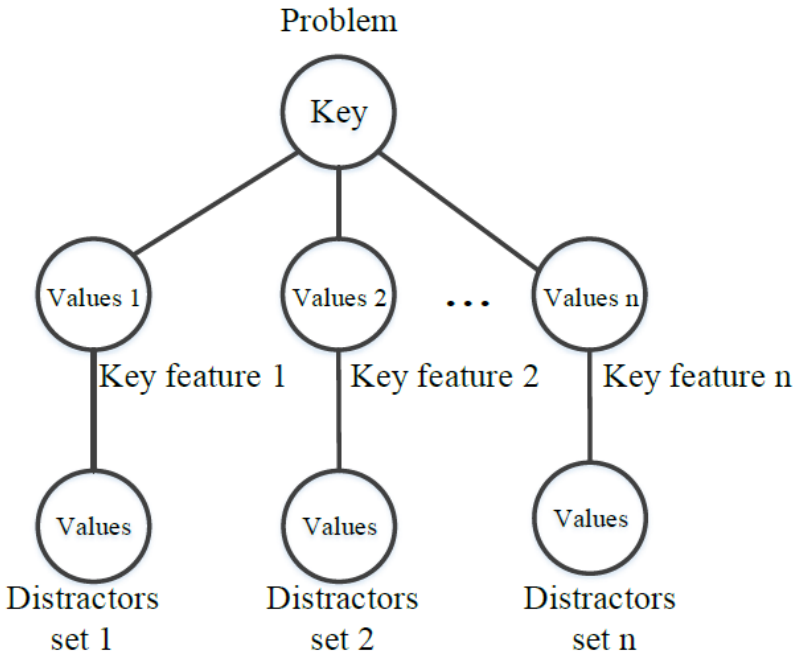


Figure 9. A complete tree for a scenario (template1).

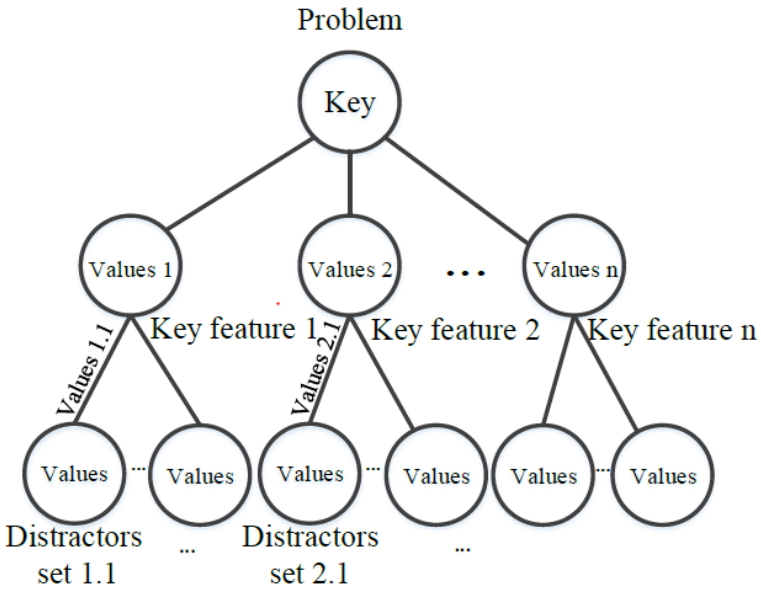


Figure 10. A complete tree for a scenario (template2).

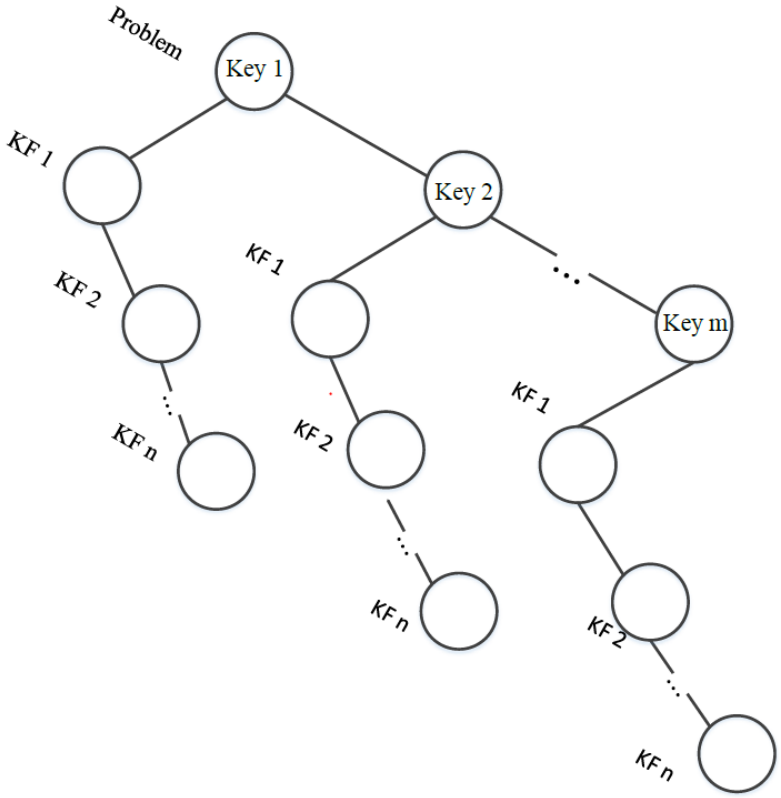


Figure 11. A binary tree for various scenarios.

Taken together, these three steps form the core requirements of content modeling. Following these steps, SMEs can choose to identify another scenario (another key), corresponding key-feature values and distractors (by returning to Step 2, Step 3 in Figure 6). This additional step usually occurs when more than one scenario exists. Thus, more than one key can be identified. SMEs can set as many keys as they need. If SMEs set multiple keys, then, for each key, there is one tree. We refer to all these trees together as a *forest*. To represent them in a complete picture, the forest will be transferred into a binary tree (Figure 10) following a specific algorithm. This binary tree may be used for content validation.

The last optional step is to identify and set additional features. Additional features are those features that do not affect how people solve the items. For example, "the hospital" of the item can be identified as an additional feature and be replaced with values, such as "ICU", "emergency room", and "clinic." Setting additional features comes as the last step because items with various keys can share all the values of additional features. Additional features increase the capacity of the model. In other words, more diverse items can be generated from the model because of the other features. This step will not affect the tree structure or the feedback model, only the item model.

By implementing these steps, the SME not only simultaneously creates a cognitive model, item model, and feedback, but also provides useful information about how to combine elements for later stages. This concept will be explained in the next section.

Content generation. Content generation is used to complete an assembly task. In this stage, variables in an item and feedback model are replaced with a permissible combination of elements to generate meaningful items and feedback (Gierl & Lai, 2016). In Gierl and Lai's framework, SMEs must complete the assembly task by programming the item model so that

constraints identified in the cognitive model can restrict implausible combinations of elements. However, programing an item model requires much effort, as users must identify constraints in the cognitive model and represent constraints in a way the generator accepts (e.g., boolean logic constraints). If feedback is required, then SMEs also need to program the feedback model, which requires more work because these constraints are not pre-defined as in the cognitive model.

To benefit users, the modified framework adopts the specific procedure described in the last section and represents the structure through a tree that records the logic underpinning the combinations of elements; hence SMEs do not need to manually set up constraints as they program models. Instead, the SMEs systematically set the key, corresponding key-feature values, distractors, and other key features in the content modeling stage. Then the logic for how to combine elements is captured by the tree, so that items and feedback can be automatically generated as the tree is traversed. In the second step (content modeling), for example, the users set the key and the corresponding values for key features. Any combination of these key features leads to the key. This relation is captured by the tree. As Figure 7 shows, any combination of the features on the second level must connect to the key. When this tree is traversed in a pre-order sequence, the elements are searched and selected in a sequence such as "problem – key feature 1 – key feature 2 … key feature n", and then logically placed into the variables of the item model to generate meaningful items. This process of capturing the logic of combined elements is also involved in step 3 (defining distractors). As the SMEs answer either the question "Can you identify distractors that do not match the following feature(s): [all values of a key feature] but match at least one of the feature(s): [the values of the rest key features]?" or the question "Can you identify distractors that do not match the following feature(s): [one value of a key feature] but match at least one of the feature(s): [the values of the rest key features]?", the information

regarding how the input distractors are related to the key is recorded. As one item is generated, the elements selected for the key features are replaced into the feedback model to form a rationale for distractors. Suppose the generated item has element A as key feature 1, and element B as key feature 2. For this item, the distractors attaching to key feature 1 will not share element A, and the distractors attaching to key feature 2 will not share element B.

Content validation. Gierl and Lai suggested validating the generated items by reviewing the cognitive model and item model. Their rationale was that if the instructions for generation in the models are sound, the outcome of the generative process should also be sound (Gierl & Lai, 2016). However, this assumption may not always be plausible because SMEs might incorrectly program the constraint logic when they input the item model into the generator. Because there is no record of these types of errors, they cannot be detected. However, with our framework, coding errors are eliminated because constraints are automatically set in the content modeling stage and then directly applied in the content generation stage. Thus, SMEs can validate the generated content by reviewing all of the models employed.

One version of the cognitive model in our framework is a forest (multiple trees). For each tree (Figure 9), the root is the key, the second level of nodes are key features, and the third level of nodes are distractors. By reviewing each tree, the users look to the key (root), to determine whether the key features (second level nodes) are plausible, and then check each of the key features (second level nodes) to determine if distractors (third level nodes) do not share these features.

An alternative version of the cognitive model is a binary tree re-constructed from the forest (multiple trees). Figure 10 presents a binary tree reconstructed from *m* trees (*m* keys). The right chain of this binary tree lists all the keys. Together this chain represents the problem, with

each node on the chain corresponding to one key. For each key, there is one chain containing the corresponding key features and their values. The nodes on these chains are arranged in the same sequence representing the same key features. However, the values inside the nodes differ. For certain key features, the nodes share all or some values. Suppose one key feature is AGE; for key 1, the corresponding age range is 1-2; for key 2, the range is 1-10; and for key 3, it is 6-10. In this way, the key feature values of key 2 cover the values of key 1 and key 3. Thus, the node for key 2 contains partially the same values as for key 1 and 3.

This binary tree presents a complete picture of the cognitive model. Like Gierl and Lai's cognitive model, it presents an abstract structure and concrete values. Unlike their cognitive model, it sets constraints automatically, through the structure of a binary tree. Recall that any combination of key features on the same chain should lead to the key connected to this chain. Furthermore, the binary tree reveals how the scenarios (keys) are related. This will provide useful information to make inferences about students' knowledge and problem-solving skills.

SMEs can also validate content by reviewing item and feedback model. The aim of this review is only to check the basic structure of items and feedback, as all necessary information has been captured by the cognitive model.

Demonstration with an Example[2]

Content modeling. Suppose the parent item is: "The color of a fruit is black, its size (diameter) is 2 cm, and it has a hard seed. Which fruit might it be? A) cherry (key), B) grape, or C) blackberry." The first step in content modeling is to identify the key information that would be used to solve this item. As the key information (black, 2 cm, and a hard seed) is identified and

[2] See corresponding screen shots in Appendix A

named as the key features (Color, Size, and Seed), an initial tree structure (Figure 12) and item model (Figure 13) are generated. This tree has three leaves associated with three key features, and the item model contains three variables.

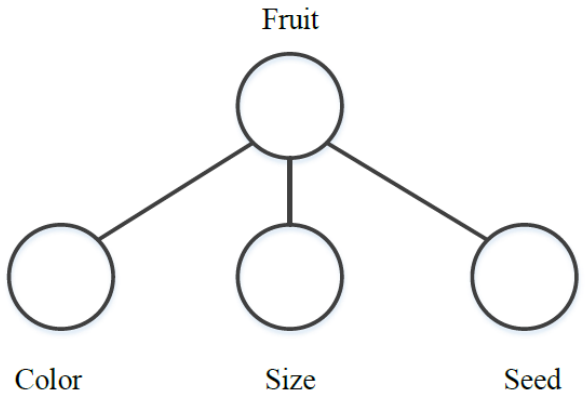


Figure 12. An initial tree.

The color of a fruit is [**color**], its size is [**size**], and it has [**seed**]. Which fruit might it be?

Figure 13. An initial item model.

The second step is to identify a scenario (key) and provide corresponding values for the variables created in step 1. Suppose the key is “cherry”. The corresponding values can be “black or red” for [color], “2cm or 3cm” for [size], “a hard seed” for [seed]. The key (“cherry”) is inserted into the root of the tree, and the values of [color], [size], and [seed] are inserted into the corresponding leaves (Figure 14). This key and these values are presented in the item model (Figure 15).

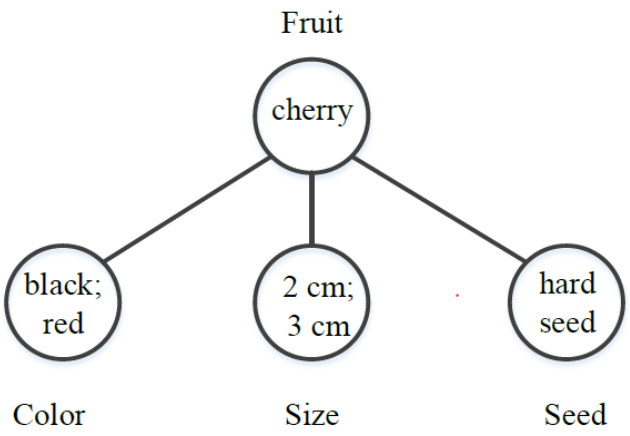


Figure 14. A tree with values.

The color of a fruit is [**color**], its size is [**size**], and it has [**seed**]. Which fruit might it be?

Options: cherry
Color: black; red
Size: 2 cm; 3cm
Seed: a hard seed

Figure 15. An extended item.

The third step is to define distractors. If the first template is used, since there are three key features, three questions are generated (see Table 7) to guide the generation of distractors. The answer to question 1 may be apricots, as apricots are neither black nor red and have a hard seed. The answer to question 2 may be roma tomatoes (RT), as roma tomatoes are neither 2 cm nor 3 cm and are red. Black plums (BP) may be an answer to question 2, as they are neither 2 cm nor 3cm, and they are black and have a hard seed. Similarly, apricots may be an answer to question 2, as they are neither 2 cm nor 3 cm and have a hard seed. The answer to question 3 may be roma tomatoes, as roma tomatoes do not have a hard seed and are red. These answers are inserted into the tree as the children of key feature leaves to show why these distractors are not plausible (Figure 16). As Figure 16 indicates, apricot is attached to [color] as implausible, because the color cannot be black or red; roma tomato (RT), black plum (BP), and apricot are attached to [size] as implausible because they cannot be 2 cm or 3 cm; roma tomato (RT) is attached to [seed] as implausible because it should not have a hard seed. These distractors are also presented in the item model (Figure 17).

If the second template is used, since there are five values of all key features, five questions are generated (see Table 8) to guide the generation of distractors. The answer to question 1 may be apricots, as apricots are not black and have a hard seed. The answer to question 2 may be apricots too, as apricots are not red and have a hard seed. Similarly, black plums may be the answer to question 2, as they are not red and have a hard seed. The answer to question 3 may be roma tomatoes (RT), black plums, and apricots, as they are not 2 cm and share certain values of [color] and [seed]. Similarly, roma tomatoes, black plums, and apricots may also be presented as answers to question 4. The answer to question 5 can be roma tomatoes, as roma tomatoes do not have a hard seed and are red. These answers are inserted into the tree as the children of key feature leaves to show why these distractors are not plausible (Figure 18). As Figure 18 indicates, apricot are attached to [black] and [red] as implausible, because the color cannot be black or red; black plum is attached to [red], because the color cannot be red; roma tomato (RT), black plum (BP), and apricot are attached to [2 cm] and [3 cm] as implausible because they cannot be 2 cm or 3 cm; roma tomato (RT) is attached to [seed] as implausible because it should not have a hard seed. Given that the two templates work similarly, we will focus for the reminder of this demonstration on using template 1 in order to avoid redundancy.

As noted above, when templates 1 or 2 are used, SMEs are required to answer three or five questions respectively. Therefore, template 1 is recommended for use when saving time is the priority. However, the extra time spent when template 2 is used brings advantages, such as diverse distractors and comprehensive feedback. For example, when template 2 is used to guide the generation of distractors that attached to the key feature “color”, SMEs can identify the distractor – black plum. This distractor can not be identified when template 1 is used, because template 1 requires distractors that are neither black nor red. As a result, in the case of the item

“The color of the fruit is color, its size is 2 cm, and it has a hard seed. Which fruit might it be?” the feedback for the incorrect choice “black plum” with template 2 is more comprehensive than with template 1. With template 1, the feedback is: this is incorrect because it does not match the following feature(s): ['2 cm'], while with template 2, the feedback is: this is incorrect because it should not appear with [‘black;'2 cm'].

Table 7

A template and three questions generated

Template: Can you identify distractors that do not match the following feature(s): [all values of a key feature] and match at least one of the feature(s): [the rest of key features]?
Three generated questions:
1.Can you identify distractors that do not match the following feature(s): [black; red] and match at least one of the feature(s): [2cm;3cm; hard seed]?
2.Can you identify distractors that do not match the following feature(s): [2 cm;3cm] and match at least one of the feature(s): [black; red; hard seed]?
3.Can you identify distractors that do not match the following feature(s): [hard seed] and match at least one of the feature(s): [black; red; 2cm;3cm]?

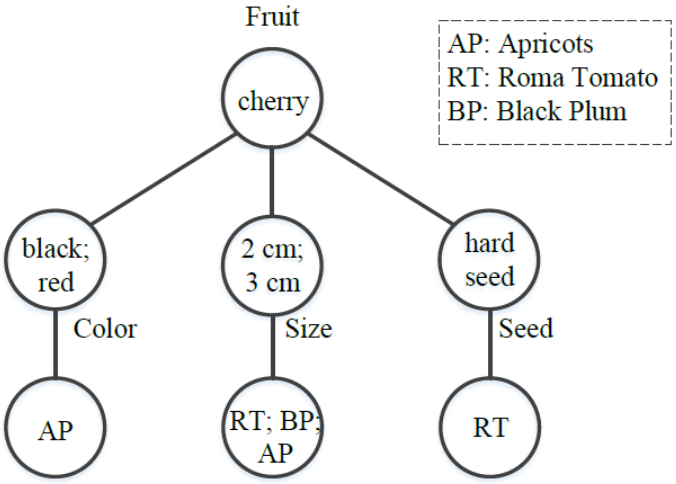


Figure 16. A tree for the key-cherry (template 1).

The color of a fruit is [**color**], its size is [**size**], and it has [**seed**]. Which fruit might it be?

Options: cherry; apricots; roma tomato; black plum; limes; avocado
Color: black; red; green
Size: 2 cm; 3cm;7 cm; 8cm
Seed: a hard seed; multiple seed

Figure 17. An extended item model.

Table 8

A template and five questions generated

Template: Can you identify distractors that do not match the following feature(s): [one value of a key feature] and match at least one of the feature(s): [the rest of key features]?

Three generated questions:

1.Can you identify distractors that do not match the following feature(s): [black] and match at least one of the feature(s): [2cm;3cm; hard seed]?

2.Can you identify distractors that do not match the following feature(s): [red] and match at least one of the feature(s): [2cm;3cm; hard seed]?

3.Can you identify distractors that do not match the following feature(s): [2 cm] and match at least one of the feature(s): [black; red; hard seed]?

4.Can you identify distractors that do not match the following feature(s): [3 cm] and match at least one of the feature(s): [black; red; hard seed]?

5.Can you identify distractors that do not match the following feature(s): [hard seed] and match at least one of the feature(s): [black; red; 2cm;3cm]?

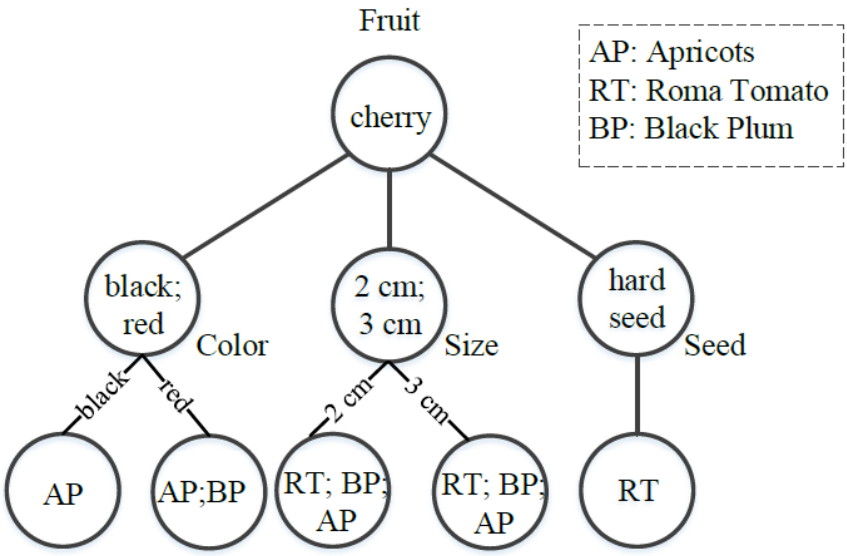


*Figure 18.*A tree for the key-cherry (template 2)

Following these three steps, another scenario (key, e.g., apple), its corresponding key-feature values, and distractors can be set. Then, a second tree (figure 19) is formed. Notice that the distractor set attached to [color] is empty. This occurs when SMEs do not identify suitable distractors. In this example, SMEs cannot come up with distractors that are either red or green and have at least one of the following features: “7cm, 8cm, and multiple seeds.” As the second tree is formed, the item model (Figure 20) must therefore be extended. The first and second tree together are referred to as a forest. To represent them in a complete picture, the forest will be transferred into a binary tree (Figure 21) following a specific algorithm (Geng, 2005). This binary tree may be used for content validation.

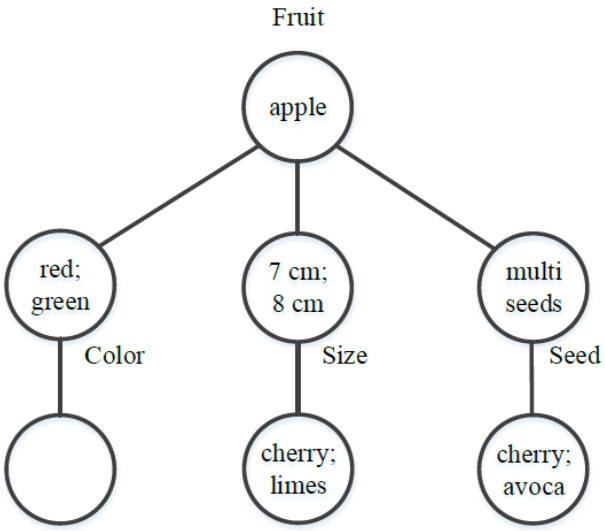


Figure 19. A tree for the key-apple.

The color of a fruit is [**color**], its size is [**size**], and it has [**seed**]. Which fruit might it be?

Options: cherry; apricots; roma tomato; black plum; limes; avocado
Color: black; red; green
Size: 2 cm; 3cm;7 cm; 8cm
Seed: a hard seed; multiple seed

Figure 20. An extended item mode.

Content generation. The elements of the content and the logic of combining these elements to generate items and feedback is captured in the trees. When each tree is traversed, the element of the root is first selected as the key, and then from each of the three key features, one element is randomly selected to replace the variables within the item model. Finally, supposing that the item has three options (one key and two distractors), two distractors are selected from the sets. Using the tree in Figure 13 as an example, when the root "cherry" is selected for the key, "black" for [color], "2 cm" for [size], "a hard seed" for [seed], "apricots" and "black plum" for distractors, the generated item becomes:

The color of fruit is black, its size is 2 cm, and it has a hard seed. Which fruit might it be?

A. Cherry

B. Apricots

C. Black plum

At the same time, the selected elements are placed into the feedback model to form rationales for each option of this specific item. Recall the feedback model for the key is defined as: [key] is

correct because it matches every feature: [all key features]. The feedback model for the distractors is: [distractor] is incorrect because it does not match the following feature(s): [certain key features]. In our example, all of the selected key features “black, 2 cm, and a hard seed” replace the [all key features]; thus, the rationale for the key becomes: cherry is correct because it matches every feature: “black, 2 cm, and a hard seed.” As the tree is back-traversed (from the leaves to its parent), we find that apricots attach to two key features: [color] and [size]. Thus, the selected elements for [color] and [size], “black” and “2 cm,” replace the [certain features]. As a result, the rational becomes: Apricot is incorrect because it does not match the following feature(s): “black” and “2 cm”. Similarly, we find that black plum attaches to [size], so the selected element for [size], “2cm,” replaces the [certain features]. Thus, the rationale is: Black plum is incorrect because it does not match the following feature(s): “2 cm”. Table 9 presents some generated items and their rationales. They clearly show that as the elements of an item change, the rationales for each option of this specific item also change.

Table 9

Examples of Generated Content

1.The color of a fruit is black, its size is 3 cm, and it has a hard seed. Which fruit might it be?
A. roma tomato: this is incorrect because it does not match the following feature(s): '3 cm', 'a hard seed'
B. apricots: this is incorrect because it does not match the following feature(s): 'black', '3 cm'
C.* cherry: this is correct because it matches every feature: 'black', '3 cm', 'a hard seed'

2. The color of the fruit is green, its size is 7 cm, and it has multiple seeds. Which fruit might it be?
A. *apple: this is correct because it matches every feature: 'green', '7 cm', 'multiple seeds'
B. avocado: this is incorrect because it does not match the following feature(s): 'multiple seeds'
C. limes: this is incorrect because it does not match the following feature(s): '7 cm'

3 The color of the fruit is black, its size is 2 cm, and it has a hard seed. Which fruit might it be?
A. black plum: this is incorrect because it does not match the following feature(s): '2 cm'
B. roma tomato: this is incorrect because it does not match the following feature(s): '2 cm', 'a hard seed'
C. *cherry: this is correct because it matches every feature: 'black', '2 cm', 'a hard seed'

4. The color of the fruit is red, its size is 7 cm, and it has multiple seeds. Which fruit might it be?
A. avocado: this is incorrect because it does not match the following feature(s): 'multiple seeds'
B. cherry: this is incorrect because it does not match the following feature(s): '7 cm', 'multiple seeds'
C. *apple: this is correct because it matches every feature: 'red,' '7 cm', 'multiple seeds'

5. The color of the fruit is red, its size is 3cm, and it has a hard seed. Which fruit might it be?
A. roma tomato: this is incorrect because it does not match the following feature(s): '3cm', 'a hard seed'
B. *cherry: this is correct because it matches every feature: 'red,' '3cm', 'a hard seed'
C. apricots: this is incorrect because it does not match the following feature(s): 'red', '3cm'

Content validation. Our framework provides two ways to validate the content. The first is to review the multiple trees (Figures 16 and 19). Based on each tree, SMEs verify whether the elements of key features are plausible. In the case highlighted in these figures, for example, they must check whether black and red are plausible values for [color] of cherry, 1 cm and 2 cm are

plausible for [size], and a hard seed is plausible for [seed]. They also must check confirm whether the identified distractors are plausible. In this case, for example, apricots are identified distractor attached to [color]. Thus, SMEs would check whether apricots are neither black nor red but have at least one of the rest of the features, such as a size of 2 cm, 3cm, or a hard seed. Similarly, roma tomatoes, black plums, and apricots are attached to [size]. Thus, SMEs must determine whether they are neither 2 cm nor 3 cm but have at least one of the rest of the features, such as a color of black, or red, or a hard seed.

The second way to review the content is through the binary tree (Figure 21). The right chain of this tree lists two keys: cherry and apple. This chain indicates that the problem involves identifying fruit. For each key, there is one chain containing corresponding key features and their values. The key features are the same: [color], [size] and [seed]. However, the values inside are different. For example, the values of [size] for cherry are 2 cm and 3 cm, as cherry is a small fruit while [size] for apple contains 7 cm and 8 cm, as apple is a medium-sized fruit. As indicated by the binary tree, for key feature [color], cherry and apple have a shared element – red.

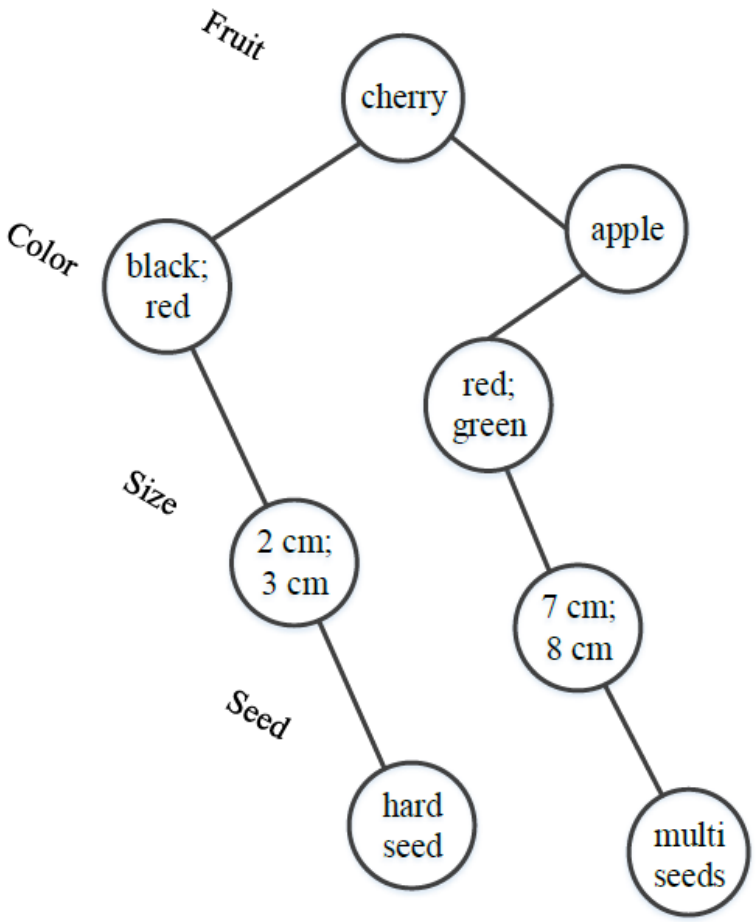


Figure 21. A binary tree for problem-fruit.

Research Design

In this chapter I propose a modified template-based framework which employs tree structures as a mechanism to create content in a more efficient way. I have designed a two-stage experiment to test the feasibility of this framework and test the evaluative and psychometric quality of the generated content.

Stage 1: Testing Feasibility. Two instructors of EDPY 303 Educational Assessment, which is a mandatory course within the teacher education program at the University of Alberta, separately used HIM_AIG (a developed software based on the modified framework) to

implement generation tasks, including content modeling, content generation, and content validation based on a parent item that is different for each instructor. Before they worked with a different parent item, each of them started with the same practice session in which step-by-step instructions for using HIM_AIG are given. As part of the first task, the instructors developed distractors with the assistance of either of the question templates (see Table 11). The content models produced, which includes cognitive models, item models, and feedback models, and content generated, which includes items and feedback were presented to demonstrate the feasibility of the framework.

There are two reasons for conducting this application within a university teacher education area. First, the instructors are faced with the challenge of creating large numbers of items to support computerized formative assessment. To help the students to become effective teachers in the future, there is a tremendous demand for computerized formative assessments of knowledge related to curriculum and instruction, classroom management, educational assessment, and the sociology of schooling (Gambhir, Broad, Evans, & Gaskell, 2008). Second, the empirical results of this research can be better disseminated in this context. The instructors of teacher education are usually the pioneers in educational reform who are more sensitive to the ongoing evolvement of educational assessment and more open to new educational trials.

Stage 2: Evaluating Quality. Six experts who hold a Ph.D. in measurement, have experience teaching assessment courses, and are not involved in the stage 1 experiment were recruited to evaluate the quality of the generated content using the rules presented in Table 10. These rules, which are selected from Table 2 and 3, achieve the greatest consensus among published item-writing and feedback-writing guidelines/recommendations. As Table 10 indicates,

one of the five rules is related to the statement, three are related to options, and one is related to feedback.

Table 10

Rules for evaluating the content quality

Statement	1.Present the main idea/intended content clearly
Options	2. Use plausible and discriminating options
	3. Keep options independent/ Avoid giving clues to the correct response
	4. Avoid using the options all of the above, none of the above, and I don't know
Feedback	5. Provide elaborated feedback (what, how, why)

As noted above, three different parent items (one used in the practice session and two used in the formal sessions) were used to generate content, and based on each parent item, one template was used to guide the development of distractors. Hence, three cognitive models (tree structures) were produced. Two items and their corresponding feedback were randomly selected from the content pool generated using each cognitive model. Presented with the three generated parent items in a single test form, the six experts rated this generated content based on the rules on a 4-point scale ranging from *strongly disagree* to *strongly agree*, where a higher score indicates more agreement that the content satisfies the rule. The overall median rate of each rule was presented to indicate the quality of the generated content.

In addition to test the evaluative quality of content generated, psychometric quality of the items generated were also be tested. The items used for testing the evaluative quality were

allocated in the field test form. Specifically, the items generated in experiment 1, experiment 2, and practice sessions were separately allocated to three different online practice quizzes, so that students' responses could be collected through the online system as they write the quizzes along the term. The students' responses were analyzed using psychometric indices from classical test theory.

Table 11

Research design

	Stage 1		Stage 2	
	Instructor 1	Instructor 2	Six experts	EDPY 303 Students
Practice Session	G	G	E	W
Experiment 1	G		E	W
Experiment 2		G	E	W

G: generating content
E: evaluating the generated content
W: writing the generated items

Chapter 4 Results

This study provides a modified framework for content generation and is intended to explore its application within the context of higher education. The purpose of the application component was achieved by conducting a two-stage experiment which tests the feasibility of this framework and evaluates the quality of the generated content. This chapter arranges the presentation of the findings by these two stages. The results of the first stage, which include cognitive models, item models, feedback models, items, and feedback, are presented first followed by the results of the second stage, which include descriptive and inferential analysis of the ratings of the generated content and the psychometric analysis of the generated items.

Testing Feasibility

Two instructors of EDPY 303 (Educational Assessment), which is a mandatory course within the teacher education program at the University of Alberta, used HIM_AIG to implement generation tasks, including content modeling, content generation, and content validation for quizzes used in the course. The instructors did so separately, using parent items that were selected to test the traditional and non-traditional application of the modified framework for content generation. The traditional application involves systematically combining related features to generate items with choices that have only one value. The non-traditional application involves systematically combining independent features to generate items with choices that have a set of values. Although each instructor worked with a different parent item, each started with the same practice session in which step-by-step instructions for using HIM_AIG were given. The results of these experiments, and of the practice sessions are presented below. (As the parent item used in the practice session is related to the non-traditional application, the corresponding results are presented in the non-traditional application section.)

Traditional application. The first experiment made use of the following parent item: "Which tool has the high objectivity in scoring, low authenticity and complexity, and low time requirement for developing? A) Performance assessment tasks B) Portfolio assessment C) Selected-response items D) Constructed-response items." Starting from this parent item, instructor A continually interacted with HIM_AIG to collaboratively create cognitive, item, and feedback models, as well as content including items and feedback. As noted in Chapter 3, two templates were designed to generate questions to guide SMEs in identifying plausible distractors. In this experiment, instructor A used template 2 because more diverse distractors and comprehensive feedback were required. Template 2 guides the instructor in identifying distractors that do not match "low time requirement for developing," "medium time requirement for developing," and "high time requirement for developing," respectively.

Figure 22 presents the cognitive model: one tree with the key "selected-response items" and the other with the key "constructed-response items." For each value of the key features, there is one set containing the distractors that do not match with the value. For example, as the left panel shows, the distractor set "Performance assessment tasks; Portfolio assessment; Constructed-response items" does not match the low time requirement. The distractor set "Performance assessment tasks; Portfolio assessment" does not match the medium time requirement. The distractor set "Constructed-response items" does not match the high time requirement. Figure 23 and 24 present the content models and some generated content, respectively.

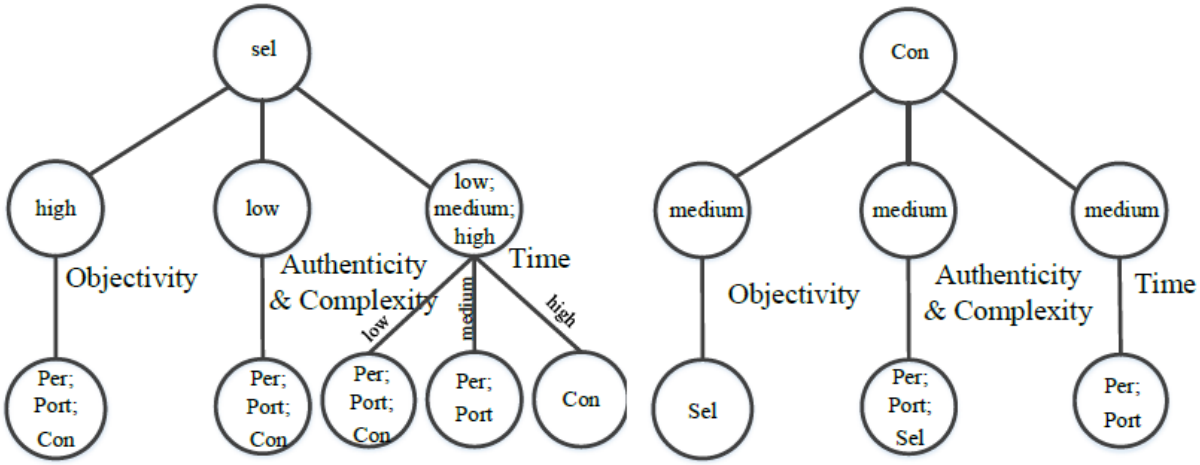


Figure 22. The tree-structured cognitive model produced in experiment 1.

Item Model:	Feedback Model:
Which tool has the [**objectivity**], [**authenticity & complexity**], and [**time**]?	[**Option**] is correct, as it matches [**all key features**]; [**Option**] is not correct because it does not match [**some key features**]
Objectivity: high objectivity, medium objectivity **Authenticity & Complexity**: low authenticity and complexity, medium authenticity and complexity **Time:** low time requirement for developing, medium time requirement for developing, high time requirement for developing **Option**: Performance assessment tasks; Portfolio assessment; Selected-response items; Constructed-response items	

Figure 23. The content models (Item and Feedback model) produced in experiment 1.

1. Which tool has the high objectivity in scoring, low authenticity and complexity, and low time requirement for developing?

    A. Performance assessment tasks: A is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, low time requirement for developing

    B. Portfolio assessment: B is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, low time requirement for developing

    C. *Selected-response items: C is correct, as it matches every feature: high objectivity in scoring, low authenticity and complexity, and low time requirement for developing

    D. Constructed-response items: D is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, low time requirement for developing

2. Which tool has the high objectivity in scoring, low authenticity and complexity, and medium time requirement for developing?

    A. Portfolio assessment: A is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, medium time requirement for developing

    B. *Selected-response items: B is correct, as it matches every feature: high objectivity in scoring, low authenticity and complexity, medium time requirement for developing

    C. Performance assessment tasks: C is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, medium time requirement for developing

    D. Constructed-response items: D is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity

Figure 24. The content (Items and Feedback) generated in experiment 1.

Non-traditional application. The second experiment made use of the parent item presented in Table 12. Unlike the parent item used in the traditional application section (experiment 1), this one presents trios as options. It presents three independent examples of student tasks in the stem and lists three corresponding Bloom's levels for these tasks in each option. In the modeling process, these independent examples are treated as three key features, and each list is treated as an option. For this experiment, template 1 was used to guide SMEs in identifying plausible distractors because it can help the instructor produce the same distractors and feedback as template 2 while requiring less work. Figure 25 presents the cognitive model： one tree with the key "Analyze; Create; Evaluate" and the other with the key "Evaluate; Analyze; Create." Figure 26 presents the item and feedback models. Figure 27 presents some generated items and feedback.

Table 12

Parent Item used in experiment 2

Example 1: "Distinguish between relevant and irrelevant numbers in a math word problem";
Example 2: "Write a research paper";
Example 3: "Check if conclusions follow the observed data"
Which of the following trios of Bloom's levels best describe examples 1, 2 and 3, respectively?
A. Analyze; Create; Evaluate
B. Analyze; Evaluate; Analyze
C. Evaluate; Create; Evaluate
D. Evaluate; Evaluate; Analyze

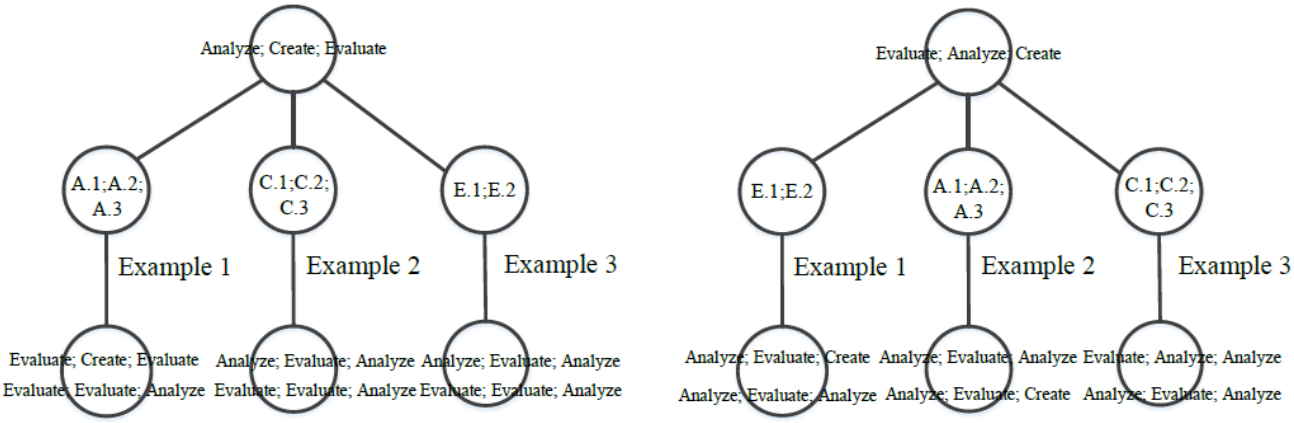


Figure 25. The tree-structured cognitive model produced in experiment 2.

Item Model:	Feedback Model:
Example 1: [**Example 1**] Example 2: [**Example 2**] Example 3: [**Example 3**] Which of the following trios of Bloom's levels best describe examples 1, 2 and 3, respectively?	[**Option**] is correct, as it matches [**all key features**]; [**Option**] is not correct because it does not match [**some key features**]**.**
Example 1, Example 2, Example 3: A.1.Distinguish between relevant and irrelevant numbers in a math word problem, A.2.Structure historical evidence for and against a particular historical explanation, A.3.Determine the point of view of an author based on his or her political perspective, C.1.Plan a research paper, C.2.Generate a hypothesis to account for observed phenomenon, C.3.Build habitats for a specific purpose, E.1.Determine if conclusions follow the observed data, E.2.Judge which of two methods is the best way to solve a problem **Option:** Analyze; Create; Evaluate/ Analyze; Evaluate; Analyze/ Evaluate; Create; Evaluate/ Evaluate; Evaluate; Analyze/ Analyze; Evaluate; Create/ Analyze; Evaluate; Create/ Evaluate; Analyze; Analyze/ Analyze; Evaluate; Analyze	

Figure 26. The content models (Item and Feedback model) produced in experiment 2.

1. Example 1: "Distinguish between relevant and irrelevant numbers in a math word problem"
   Example 2: "Build habitats for a specific purpose"
   Example 3: "Check if conclusions follow the observed data"
   Which of the following trios of Bloom's levels best describes examples 1, 2 and 3, respectively?

   A. Evaluate; Create; Evaluate: A is incorrect, because it does not match the following example(s): “Distinguish between relevant and irrelevant numbers in a math word problem"

   B. *Analyze; Create; Evaluate: B is correct, as it matches every example: "Distinguish between relevant and irrelevant numbers in a math word problem", "Build habitats for a specific purpose”, "Check if conclusions follow the observed data"

   C. Evaluate; Evaluate; Analyze: C is incorrect, because it does not match the following example(s): "Distinguish between relevant and irrelevant numbers in a math word problem", "Build habitats for a specific purpose”, "Check if conclusions follow the observed data"

   D. Analyze; Evaluate; Analyze: D is incorrect, because it does not match the following example(s): "Build habitats for a specific purpose”, "Check if conclusions follow the observed data"

2. Example 1: "Judge which of two methods is the best way to solve a problem "
   Example 2: "Structure historical evidence for and against a particular historical explanation"
   Example 3: "Generate a hypothesis to account for observed phenomenon"
   Which of the following trios of Bloom's levels best describes examples 1, 2 and 3, respectively?

   A. *Evaluate; Analyze; Create: A is correct, because it matches every example: "Judge which of two methods is the best way to solve a problem ", "Structure historical evidence for and against a particular historical explanation”, "Generate a hypothesis to account for observed phenomenon"

   B. Analyze; Evaluate; Create: B is incorrect, because it does not match the following example(s): "Judge which of two methods is the best way to solve a problem ", "Structure historical evidence for and against a particular historical explanation”

   C. Evaluate; Analyze; Analyze: C is incorrect, because it does not match the following example(s): “Generate a hypothesis to account for observed phenomenon”

   D. Analyze; Evaluate; Analyze: D is incorrect, because it does not match the following example(s): “Judge which of two methods is the best way to solve a problem ", "Structure historical evidence for and against a particular historical explanation”, "Generate a hypothesis to account for observed phenomenon"

Figure 27. The content (Items and Feedback) generated in experiment 2.

The parent item used in the practice session was: “Which of the following pairs best describe the levels at which "difficulty" and “reliability" are analyzed, respectively? A) test level; test level; B) item level; test level; C) test level; item level; D) item level; item level.” Like the parent item used in experiment 2, this item presents independent instances in the stem as key features and lists corresponding information for each element as options. Figure 28 presents the cognitive model： one tree with the key “item level; test level” and the other with the key “test level; item level.” Figure 29 presents the item and feedback models. Figure 30 presents some generated items and feedback.

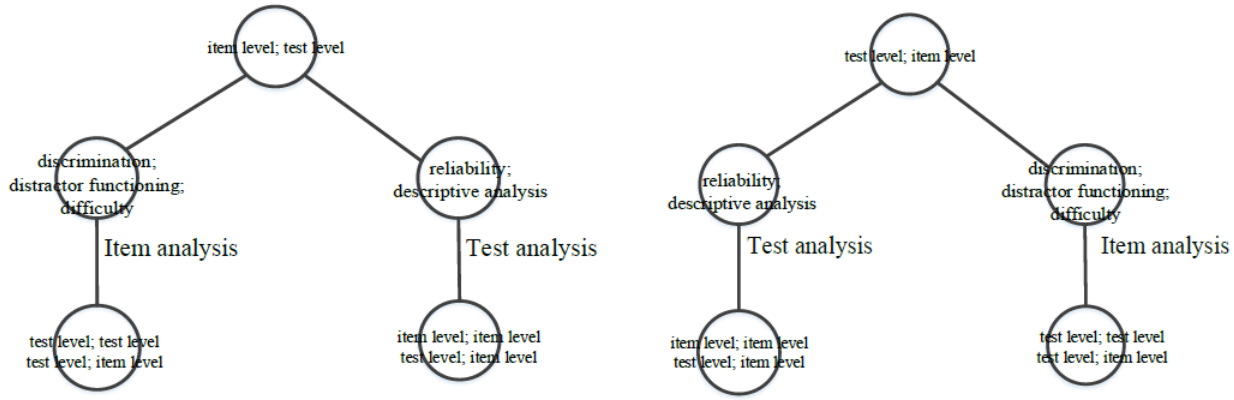


Figure 28. The tree-structured cognitive model produced in practice sessions.

Item Model:	Feedback Model:
Which of the following pairs best describe the levels at which "[**Analysis 1**]" and "[**Analysis 2**]" are analyzed, respectively?	[**Option**] is correct, as it matches [**all key features**]; [**Option**] is not correct because it does not match [**some key features**].
Analysis 1, Analysis 2: difficulty, discrimination, distractor functioning, reliability, descriptive analysis of test scores **Option:** item level; test level/ test level; test level; item level; item level/test level; item level	

Figure 29. The content models (Item and Feedback model) produced in the practice session.

1. Which of the following pairs best describe the levels at which "descriptive analysis of test scores" and “discrimination" are analyzed, respectively?

   A. test level; test level: A is incorrect, because it does not match the following term(s): “discrimination"

   B. item level; test level: B is incorrect, because it does not match the following term(s): "descriptive analysis of test scores", “discrimination"

   C. *test level; item level: C is correct, as it matches every term: "descriptive analysis of test scores", “discrimination"

   D. item level; item level: D is incorrect, because it does not match the following term(s): "descriptive analysis of test scores"

2. Which of the following pairs best describe the levels at which "descriptive analysis of test scores" and “distractor functioning" are analyzed, respectively?

   A. test level; test level: A is incorrect, because it does not match the following term(s): “distractor functioning"

   B. item level; item level: B is incorrect, because it does not match the following term(s): "descriptive analysis of test scores"

   C. item level; test level: C is incorrect, because it does not match the following term(s): "descriptive analysis of test scores", “distractor functioning"

   D. *test level; item level: D is correct, because it matches every term: "descriptive analysis of test scores", “distractor functioning"

Figure 30. The content (Items and Feedback) generated in the practice session.

Testing Quality

Evaluative Quality. Six experts who hold a Ph.D. in measurement and have an average of 4.5 years of experience teaching assessment courses were recruited to evaluate the quality of the generated content using the statement presented in Table 12 (see the complete survey in Appendix B). These statements were derived from the rules presented in Table 10. Statement 1 ("The item presents the main idea/intended content clearly") is a duplicate of rule 1; statement 2 ("The item provides plausible options") and another statement, "the item provides discriminating options" are extracted separately from rule 2 to avoid conveying double-barrelled information. However, as the discrimination function will be assessed through psychometric analysis, the statement "the item provides discriminating options" is not presented in Table 12. Statement 3 ("The item provides independent options which give no clues to the correct response") is paraphrased from rule 3; statement 4 ("The feedback provides an elaborated explanation of the item") matches rule 5. Rule 4 ("Avoid using the options all of the above, none of the above, and I don't know") is not covered by these statements, because this rule was automatically met as none of the options contain "all of the above"," none of the above" or "I don't know".

Table 13

Statements used for evaluating the quality of the generated content

1. The item presents the main idea/intended content clearly.
2. The item provides plausible options.
3. The item provides independent options which give no clues to the correct response.
4. The feedback provides an elaborated explanation of the item.

As noted above, three different parent items (two non-traditional and one traditional ones) were used to generate content, and three cognitive models (tree structures) were produced. From the content pool generated using each cognitive model, one parent item and its corresponding feedback, and two randomly selected non-parent items and their corresponding feedback were presented in a single test form (In total, nine items and their corresponding feedback). The parent content was used as a point of reference for determining the quality of the non-parent content. If the modified framework is feasible, then the quality of the non-parent content should not be worse than the parent content. The experts evaluated their quality based on the statements on a 4-point scale ranging from *strongly disagree* to *strongly agree*, where a higher score indicates more agreement that the content satisfies the statement. To assess inter-rater reliability, intraclass correlation (ICC) was computed through an absolute-agreement, 2-way mixed-effects model. The resulting ICC value was 0.61, indicating moderate reliability among the six experts (Koo & Li, 2016).

Table 13 and 14 present the median ratings of each statement for the parent and non-parent content, respectively. The descriptive statistics show the following results.

1) Both parent and non-parent items generated met the recommended quality standard, as the median ratings of statement 1 to 3, which are related to the items, are either 3 or 3.5 (3 means “agree with the statement”).
2) However, the feedback generated did not do so, as the median ratings of statement 4, which is related to the feedback, are 2 (2 means “disagree with the statement”).

Table 14

Median ratings for content quality statements of parent content

Rule	Model 1	Model 2	Model 3	Overall
1	3	3.5	3	3
2	3	3.5	3	3
3	3	3	3	3
4	2	2	2.5	2

Table 15

Median rating for content quality statements of non-parent content

Rule	Model 1	Model 2	Model 3	Overall
1	3	3	3	3
2	3	3	3	3
3	3	3.5	3	3
4	2	3	2	2

The analysis of open-ended survey questions reveals that the main concern of the experts who rated the quality of feedback as low was that the feedback failed to provide sufficient explanations of the incorrect options. Related comments are presented below.

"I hope the feedback can be more educational rather than simply pointing out where students did wrong."

"The feedback could be more structured to identify the origin of the misconception just rather than pointing out the incorrect components."

"I found that most of the elaborated feedback only indicated which portion of the question was incorrect but does not explain why."

Recall that the parent content was used as a point of reference for determining the quality of the non-parent content. To test whether the quality of the non-parent items is statistically higher or equal to the quality of parent ones, four Mann-Whitney U Tests were conducted using SPSS. For each test, the dependent variable was the median rating for each of the statements and the independent variable was the content type. The results show that the ratings regarding statement 1("The item presents the main idea/intended content clearly") for non-parent items (Mean Rank = 26.71) did not differ significantly from parent items (Mean Rank = 27.56), U = 305, z = −.207, p = .836; ratings regarding statement 2 ("The item provides plausible options") for non-parent items (Mean Rank = 26.85) did not differ significantly from parent items (Mean Rank = 25.83), U = 318, z = .266, p = .790; ratings regarding statement 3 ("The item provides independent options which give no clues to the correct response") for non-parent items (Mean Rank = 27.63) did not differ significantly from parent items (Mean Rank = 25.78), U = 337, z = .442, p = .658; ratings regarding statement 4 ("The feedback provides an elaborated explanation of the item") for non-parent items (Mean Rank = 27.16) did not differ significantly from parent items (Mean Rank = 26.69), U = 320.5 z = .112, p = .911.

These non-significant results indicate that non-parent content has comparable quality with that of the parent content. They further imply that the modified framework is feasible. As a result, the low quality of feedback should not be attributed to the framework itself, but from the low quality of the parent feedback or the design of feedback model. This point is further discussed in Chapter 5.

Psychometric Quality. According to the design presented in Chapter 3, the field test sample should have included the nine generated items used in the Evaluating Quality Section. Specifically, the items generated in experiment 1, experiment 2, and practice sessions should

have been separately allocated to three different online practice quizzes by the TA of EDPY 303 who was responsible for assembling items into online quizzes, so that students' responses could have been collected through the online system as they wrote the quizzes along the term. However, only the three items generated within experiment 2 were tested because only these items were allocated to an online quiz. The others were inadvertently overlooked by the TA. This administrative error was not found until EDPY 303 ended when collecting additional data was impossible. Hence, only the results for experiment 2 are reported.

The number of valid students' responses was 290. Item analyses were conducted using psychometric indices from classical test theory. Item difficulty and discrimination were computed. Item difficulty was assessed via the proportion of examinees correctly answering each item. Discrimination was assessed via point-biserial correlation between examinees' responses to the choice and their totals scores, and index of discrimination (the difference between the proportion of upper 25% group who selected the choice and the proportion of lower 25% group). Column 1 of Table 15 presents the item difficulty for each generated item. Column 2 and 3 present the biserial correlation and the index of discrimination respectively for each option.

Table 16

Classical item analysis results for the options across the generated items for experiment 2

Generated Item	Options	Difficulty	Point- Biserial correlation	Index of discrimination
1 (parent)	A*	0.86	0.27	0.19
	b		-0.16	-0.06
	c		-0.11	-0.08
	d		-0.13	-0.05
2	a		-0.15	-0.10
	B*	0.85	0.33	0.19
	c		-0.17	-0.03
	d		-0.19	-0.05
3	a		-0.25	-0.12
	B*	0.86	0.37	0.17
	c		-0.09	-0.02
	d		-0.12	-0.02

*: correct options

The item difficulty values were high and stable, indicating that the three generated items were equally easy. The point-biserial correlation values of the correct options ranged from a low of 0.27 to a high of 0.37 (Mean=0.32, SD=0.04). The index of discrimination values of the correct options ranged from 0.17 to 0.19 (Mean=0.18, SD=0.01). These results regarding discrimination indicate that the generated items differentiated among examinees of different ability levels.

For the incorrect options, the point-biserial correlation values ranged from a low of -0.25 to a high of -0.09. The mean biserial correlation across the nine options was -0.15, with a standard deviation of 0.05. The index of discrimination values of the incorrect options ranged from -0.12 to -0.02. The mean index of discrimination across the nine options was -0.15, with a standard deviation of 0.05. These results indicate that the distractors differentiated low from high performing examinees.

The comparison of the item analysis results between the correct options and incorrect options reveals that the point-biserial correlation was positive for the correct options whereas the point-biserial correlation for the distractors was in all cases negative. The index of discrimination was largest and positive for the correct option whereas for the distractors they were in all cases negative. This again confirms that the distractors differentiated low from high performing examinees.

Summary

By qualitatively presenting the content models produced and the content generated, we concluded that the proposed framework was feasible. By quantitatively analyzing the objective ratings from the six experts, we concluded that while the quality of the items generated met the standard, the quality of the feedback generated needed to be improved. By quantitatively analyzing the students' responses, we assured the psychometric quality of the generated items. The next chapter discusses these major findings and to expand upon the concepts that were studied to provide insights regarding practice.

Chapter 5 Discussion

Chapter 5 restates the purpose of the study, discusses the major findings and expands upon the concepts that were studied to provide insights regarding practice. It also identifies the limitations and presents suggestions for future research on links between different types of knowledge and designs of feedback model, on hybrid AIG methodology, and on difficulty modeling through tree-structure cognitive models.

Purpose of the Study

The twofold purpose of this study was: 1) proposing a modified framework to address the operational difficulties that arise when applying template-based AIG and 2) exploring the application of this modified framework for item generation within the context of higher education. The proposed framework adopted a human-machine interactive approach and employed tree structure as a cognitive model, a mechanism to assemble elements, and a validation tool. It was applied within the university teacher education field. Specifically, two instructors of EDPY 303 (Educational Assessment) used HIM_AIG, which is a software developed based on the modified framework, to implement generation tasks for quizzes used in the course. Six experts in higher education evaluated the generated content using the statements that have achieved the greatest consensus among published item-writing and feedback-writing guidelines/recommendations. One hundred and thirty-four students took the practice quizzes, which contained the generated items.

Discussion of the Findings

The last two decades have witnessed important developments in the area of AIG research. The theoretical concepts and the practical considerations of template-based AIG for educational assessment are well documented. Within the literature, the most comprehensive work has been done by Gierl and Lai. As noted in Chapter 2, their past work (Gierl & Lai, 2013, Gierl & Lai, 2016b, Gierl & Lai, 2016c) formed a complete automatic content generation framework, which comprised three stages: content modeling, content generation, and content validation. The research questions explored by their studies can be categorized as being related to feasibility and quality. This study presents a synthesized modified version of their past work. In particular, it presents a three-stage framework which employs tree structure to achieve key goals related to each of these stages. This section discusses the findings and the implications for practice.

Research Question One

What is the feasibility of the modified framework?

The feasibility of the modified framework was demonstrated in the three operational sessions in which the instructors of EDPY 303 created content following the stages of the modified framework. During the implementation, content models were created without requiring extra inductive work from the instructor; items and feedback for individual options were automatically generated without the need for programming; and a comprehensive structure for content validation was efficiently provided. The experiment, therefore, met the expected goals, ensuring the feasibility of the modified framework for actual use.

Some researchers might critique this framework by claiming that it involves humans, and that it might therefore be less efficient than a non-template-based method which involves no humans. Both frameworks, however, are efficient; each involves humans. The misconception

that a non-template-based framework does not do so originates from the fact that the human-machine interaction is not obvious. A non-template-based framework shifts the location of the interaction between humans and machines from the front end to the back end. For example, within a sequence-based approach (see Chapter 2), which represents one category of the non-template-based framework, humans are involved in the back end where they train and tune data for algorithms to correct inaccuracies in machine predictions.

In the template-based framework put forward by this study, the human-machine interaction occurs at the front end. It enables the SMEs to supervise and control each stage of the generation task, ranging from selecting key features to setting up other features. The SMEs can also adjust the question template to set up distractors according to their needs. For example, in the traditional experiment, the SMEs used template 2 because more diverse distractors and comprehensive feedback were required. In the non-traditional experiment, the SMEs used template 1 because it helped the instructor produce the same distractors and feedback as template 2 but required less work.

The workflow of the non-template framework reflects the commonly accepted view that item writing is both an art involving subjective design and judgments, and a science involving objective rules. In practice, in order to produce high-quality items and feedback, SMEs should continually improve their skills from both angles.

Research Question Two

What is the evaluative quality of the items and feedback generated with the modified framework?

The findings related to the descriptive analysis of the ratings indicate that while the generated items met the recommended quality standard, the generated feedback did not do so. The analysis of open-ended survey questions reveals that the main concern of the experts who

rated the quality of feedback as low was that the feedback failed to provide sufficient explanations of the incorrect options.

The feedback was generated from the default feedback template: " [distractor] is incorrect, because it does not match the following feature(s):[certain key features]." This feedback template was modified from the assured distractor rationale presented in Gierl and Lai's research (2018). Within their "Key Features Set with Distractor Rationale" (as described in Chapter 2), they explain why the distractors yield the incorrect solution by using the template "The incorrect options are not plausible because [Distractor Rationale]." The values for the variable "Distractor Rationale" include the key features that the incorrect options should not match. Although their generated feedback was not what our experts called for, interestingly, its quality was thought to satisfy the standard.

The discrepancy between the quality of Gierl and Lai's rationale models and the quality of the default feedback model in this study may be explained by the nature of the questions and the degree to which information is synthesized. The questions used in Gierl and Lai's research are related to disease diagnosis in medical education. They aim to test whether students can remember the symptoms for a diagnosis. Therefore, pointing out the key features (symptoms) that the incorrect options should not match can give sufficient explanation for why selecting the choice is wrong. However, for the questions used in this study, doing so might not be sufficient, because these questions test students' comprehension ability which requires a higher-level cognition than memorization. Furthermore, Gierl and Lai provided feedback for all three incorrect options by synthesizing the information related to each answer. In the current study, feedback was provided for individual distractors without the need to synthesize information and

therefore experts might rate the quality of the former as high because they reflect a higher degree of synthesis.

The implication for practice is in order to achieve high-quality feedback, specific consideration should be given to a particular discipline when designing feedback models. These feedback models should have a high quality, as the quality of models determines the quality of the generated content. This was supported by the fact that Wilcoxon signed-rank tests were not significant. In the case of the item models derived from good quality parent items, the generated items met the recommended quality standard. In the case of the default feedback model, the generated feedback could not provide sufficient explanation for the incorrect options.

Research Question Three

What is the psychometric quality of the items generated with the modified framework?

The results of the item analysis ensured the psychometric quality of the items generated within experiment 2. However, due to the incomplete collection of students' responses, the psychometric quality of the items generated within experiment 1 and the practice sessions is unknown. Nonetheless, these limited results provide some interesting insights.

In this study, the generated items had stable difficulty levels. This contrasts to the findings of Gierl and Lai that items generated in their study "in each content area displayed a variety of difficulty levels ranging from very easy to very difficult" (Gierl & Lai, 2016, p.208). Because of this, some researchers might critique the quality of the modified framework, claiming it can only yield cloned test items performing similarly in the item analysis. However, the items generated in this study were not cloned: If they had been, then the responses provided by each student would be consistent, and each student would have answered all of the three items either correctly or wrongly. Table 17 shows that students' responses were not consistent. Students were

distributed across the eight response patterns. Furthermore, the true response patterns should be more inconsistent than the one presented in Table 17, because the potential carry-over effects might have caused students to provide consistent responses (recall that the three generated items were allocated into the same quiz). All of these results indicated that the items measured multiple facets of students' knowledge within Bloom's taxonomy.

Table 17

The frequencies of students' response patterns to the three items generated in experiment 2

Response Pattern	Frequency	Percent
111	202	69.7
110	27	9.3
101	14	4.8
100	6	2.1
011	15	5.2
010	2	0.7
001	17	5.9
000	7	2.4
Total	290	100

1: answer correctly; 0: answer incorrectly

This phenomenon of stable difficulty levels might be explained by the small size of the test pool, the narrow breadth of content covered by the items, or the within-group design used for response collection. In experiment 2, the test pool for each content area consisted of three generated items, while in Gierl and Lai's research the pool consisted of nine generated items. This small size of the test pool might result in stable difficulty levels. As more generated items are incorporated into the test pool, the difficulty levels of these items might be expected to fluctuate.

The stable difficulty levels might also be explained by the narrow breadth of content covered by items generated in experiment 2. In experiment 2, the cognitive model measured students' knowledge of Bloom's levels through only one type of variable ("examples"), while

Gierl and Lai's cognitive model measured their knowledge of the causes of neonatal jaundice through multiple variables, including feeding patterns, physical exam characteristics, and laboratory results. The more types of key features that a cognitive model contains, the wider the breadth of content covered by the generated items. As a result, the breadth of content covered by the generated items in experiment 2 was narrow, which contributed to the stable difficulty levels.

The within-group design used for collecting responses might provide a third explanation for the stable difficulties. In experiment 2, the AIG items were distributed within a single form, while in Gierl and Lai's study, a mixed design was used, where the AIG items were distributed within and across field test examination forms. Because the mixed design cannot remove individual differences completely as the within-group design does, the remaining individual differences might inflate the actual difficulty levels. Therefore, the difficulty levels in experiment 2 became relatively stable.

Taken together, these findings suggest that accurately identifying key features and systematically combining them (as the modified framework required) are essential to generate items that look parallel but measure diverse aspects of knowledge.

Limitations

Issues related to the theoretical framework. The modified framework relies on two assumptions. The first is that individualized feedback (a rationale for each option) works better than aggregated feedback (a rationale for all options). The second is that the SMEs are able to select the appropriate template for setting up distractors. These two assumptions lead to the limitations of this study.

No evidence in fact exists about whether individualized or aggregated feedback is more useful. Individualized feedback was considered more useful in this study because it can better

help students learn in a computerized formative assessment environment, specifically when parallel items are assigned to a student through a digital tool, such as iPad or phone. Individualized feedback can help the student quickly focus on the key information because it is usually more concise than the aggregated feedback, thus being more readable. Moreover, through short-term memory, these short pieces of feedback can better assist students in acquiring new knowledge. Individualized feedback is also practical because unlike aggregated feedback, it does not require that the distractors share common elements. Aggregated feedback also has its own unique advantages, such as presenting a relatively complete interpretation of the distractors, and helping students understand the problem at a synthesized level. More evidence therefore is needed to ensure that the assumption – that individualized feedback (a rationale for each option) works better than the aggregated one – is accurate.

SMEs were also assumed to be able to select the appropriate template for setting up distractors. However, more training may be required to help them understand the differences between the question templates and to help them link the appropriate one to their practical needs. In the practice sessions, both of the instructors were given verbal instructions on how to select the appropriate template. In the experiments, however, while one instructor independently completed the experiment, another asked again for instructions regarding which template should be used to set up the distractors. Future study might focus on simplifying this process through a decision system to support SMEs selecting the most appropriate question template.

Issues related to the application. One of the limitations related to the application is that the psychometric data collection did not occur as expected because of the administrative error noted in Chapter 4. As a result, fewer generated items than expected were tested. This might mean our conclusion regarding the psychometric quality of generated items is not be

generalizable. Moreover, other potential findings might be ignored, as the probability of observing new findings is usually positively related to the size of the tested item pool.

Another limitation is that the experiments were only conducted within a teacher education area. This specific area was chosen because the instructors are faced with the salient challenge of creating large numbers of items to support computerized formative assessment and because the empirical results of this research can be better disseminated. However, this affects the generalizability of the modified framework to some degree. In order to assure this generalizability, experiments conducted in different areas will be needed.

When collecting the psychometric data, students' responses were assumed to be independent. However, they were in some cases dependent, as students repeatedly took the same practice quiz at different times. This assumption might have inflated the difficulty levels of the items and the consistent levels of response patterns. In other words, the generated items might have been easier, and the students' response patterns, in particular, the patterns "111", "110", "011" or "101" (indicating that students answered all or most of the items correct) might have appeared more frequently. Even so, the conclusion made regarding the quality of generated items – the psychometric quality was ensured – was still trustworthy (please refer to the discussion of research question 3 for details).

Recommendations for Further Research

There are at least three directions for future research. First, future studies should focus on the links between different types of knowledge and the design of the feedback model. As discussed above, pointing out the key features that distractors do not match successfully explained why the distractors were wrong for items measuring declarative knowledge, while this same type of feedback failed to do so for items measuring conceptual knowledge. This implies

potential links between the types of knowledge and the designs of feedback model. Future research is needed to further explore these links. For example, potential research questions may be: what are the different designs of feedback models for items measuring different knowledge? and How are these designs implemented? If solution paths are determined to be the most powerful design in a feedback model measuring procedural knowledge, then the next question to explore is how to generate such paths automatically. Answers to these questions can facilitate the design and development of automated feedback to support learning and help create a set of guidelines relating to formative feedback.

Second, future studies might focus on interdisciplinary research on a hybrid AIG framework. As pointed out in Chapter 2, education researchers overcome pedagogical challenges with template-based approaches, but they rarely acknowledge the existence of non-template-based approaches. Conversely, researchers mainly from the linguistic computing area have actively developed computationally advanced non-template-based approaches for AIG, but they rarely consider its pedagogical uses (Zhang & Gierl, 2018).

Interdisciplinary research on hybrid AIG frameworks can bridge this gap between education scholars who focus on template-based approaches and those in the linguistic computing field who focus mainly on non-template-based approaches. Moreover, this interdisciplinary research can achieve the goals that the single form cannot achieve because they can combine the advantages of both frameworks This is true, for example, when generating English items which require the examinees to choose the correct option to fill the gap in a sentence. The statement and questions can be generated in a non-template-based approach to ensure the accuracy of the sentences, while the options can be generated using a semantic-based

approach, which automatically provides semantically close concepts as distractors, to avoid manually setting distractors.

Third, future studies might focus on using tree-structure cognitive models for difficulty modeling. Difficulty modeling is a process that relates item features to an item's psychometric characteristics (e.g., difficulty, discrimination). It is an economical solution because it can help accurately predict the psychometric characteristics of an item in advance so that the SMEs do not need to calibrate each item individually. Although a few studies (Choi, 2017; Graf & Fife, 2013; Enright & Sheehan, 2002) have claimed that difficulty modeling has the potential to support automatic generation of items with known psychometric characteristics, very little research has been undertaken in this area. A common explanation is that no psychometric methods are available. However, various methods have emerged during the last two decades, such as the classic regression model, tree-based regression model, and hierarchical regression model. The real reason might be that it is too complicated to quantify the item features. For example, no research to the best of my knowledge has successfully quantified the composition of the distractor set, which is an important predictor of difficulty levels, because the relationships between each distractor and the key, and the relationships among the distractors are hard to depict.

The tree structure put forward in this study can solve this problem, because it can capture the relationships between the key and each distractor and those among the distractors. For example the tree in Figure 22 shows that the distractor "portfolio assessment" does not share two key features "medium authenticity and complexity", and "medium time requirement for developing" with the key " constructed response item", and the distractor " selected response" does not share two key features "medium objectivity in scoring", and "medium authenticity and

complexity" with the key; and these two distractors have in common that neither of them share the key feature "medium authenticity and complexity". This information can be used to quantify the item features. Once the item features can be modeled, difficulty modeling becomes straightforward. Furthermore, the tree structure can provide a mechanism to systematically vary the features of interest by manipulating some variables while holding others constant. This can disentangle the features that contribute to the psychometrics, thus providing sound evidence for difficulty modeling.

Conclusion

Because they offer immediate feedback and on-demand testing, computerized formative assessment is becoming popular in post-secondary education. This has created a pressing need for large numbers of content-specified items together with their feedback so that the items may be available continuously for formative feedback. To support CFA in post-secondary education, I proposed a modified generation framework that was designed to help overcome the limitations of the existing frameworks. Specifically, the proposed framework adopts a human-machine interactive approach and employs tree structure – one type of the data structure – as a triple-function engine including a cognitive model, a mechanism to assemble elements, and a validation tool. This is the first time that the three stages (modeling, generation, and validation) of the template-based framework are implemented through a single data structure. The results suggest that this tree-based approach is feasible; the evaluative quality of the generated items and feedback are comparable with that of the parent ones; and the psychometric quality of the generated items satisfy the standard. Future research should focus on links between different types of knowledge and designs of feedback model, on hybrid AIG methodology, and on difficulty modeling through tree-structure cognitive models.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). Standards for educational and psychological testing. Washington, DC: American Educational Research Association.

Albano A.D., Rodriguez M.C. (2018) Item Development Research and Practice. In: Elliott S., Kettler R., Beddow P., Kurz A. (eds) Handbook of Accessible Instruction and Testing Practices. Springer, Cham https://doi.org/10.1007/978-3-319-71126-3_12

Ali, H., Chali, Y., & Hasan, S. A. (2010, June). Automation of question generation from sentences. In *Proceedings of QG2010: The Third Workshop on Question Generation* (pp. 58-67).

Arendasy, M. E., Sommer, M., & Mayr, F. (2012). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology*, *43*(3), 464-479.

Baghaee, T. (2017). *Automatic Neural Question Generation using Community-based Question Answering Systems* (Doctoral dissertation, Lethbridge, Alta.: University of Lethbridge, Dept. of Mathematics and Computer Sciences).

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., & Revuelta, J. (2002). A feasibility study of on-the-fly item generation in adaptive testing. *ETS Research Report Series*, *2002*(2), i-44.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5-25.

Bikel, D. M., Schwartz, R., & Weischedel, R. M. (1999). An algorithm that learns what's in a name. *Machine learning*, *34*(1-3), 211-231.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: principles, policy & practice*, *5*(1), 7-74.

Bormuth, J. R. (1970). *On the Theory of Achievement Test Items: With an Appendix" On the Linguistic Bases of the Theory of Writing Items", by P. Menzel*. University of Chicago Press.

Briscoe, T. (2011). Introduction to Linguistics for Natural Language Processing. *Computer Laboratory, University of Cambridge.October*, *4*.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005, October). Automatic question generation for vocabulary assessment. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 819-826). Association for Computational Linguistics.

Burrow, M., Evdorides, H., Hallam, B., & Freer-Hewish, R. (2005). Developing formative assessments for postgraduate students in engineering. *European Journal of Engineering Education*, *30*(2), 255-263.

Canadian Association of University Teachers (2018). Retrieved from https://www.caut.ca/resources/almanac/3-students

Cameron, B., & Dwyer, F. (2005). The effect of online gaming, cognition and feedback type in facilitating delayed achievement of different learning objectives. *Journal of Interactive Learning Research*, *16*(3), 243-258.

Clariana, R. B., & Lee, D. (2001). The effects of recognition and recall study tasks with feedback in a computer-based vocabulary lesson. *Educational Technology Research and Development*, *49*(3), 23-36.

Chen, W., Aist, G., and Mostow, J. (2009). *Generating questions automatically from information text*. In Proceedings of AIED 2009 Workshop on Question Generation, pages 17–24.

Choi, J**.** (2017). Next Generation Item and Test Development: A Practical Introduction to Automatic Item Generation. Washington, D.C.: Assessment, Testing and Measurement Technical Report Series, The George Washington University.

Collins, M., Carnine, D., & Gersten, R. (1987). Elaborated corrective feedback and the acquisition of reasoning skills: A study of computer-assisted instruction. *Exceptional Children*, *54*, 254–262.

Connor, C. M., Morrison, F. J., Fishman, B., Crowe, E. C., Al Otaiba, S., & Schatschneider, C. (2013). A longitudinal cluster-randomized controlled study on the accumulating effects of individualized literacy instruction on students' reading from first through third grade. *Psychological Science*, *24*(8), 1408-1419.

Corbett, A. T., & Anderson, J. R. (2001, March). Locus of feedback control in computer-based tutoring: Impact on learning rate, achievement and attitudes. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 245-252). ACM.

Corbalan, G., Paas, F., & Cuypers, H. (2010). Computer-based feedback in linear algebra: Effects on transfer performance and motivation. *Computers and Education*, *55*, 692–703.

Cox, K., Imrie, B. W., & Miller, A. (2014). *Student assessment in higher education: a handbook for assessing performance*. Routledge.

Danon, G., & Last, M. (2017). A Syntactic Approach to Domain-Specific Automatic Question Generation. *arXiv preprint arXiv:1712.09827*.

Dearling, Ron (1997). *Higher Education in the Learning Society,* HMSO, London.

Dennis, I., Handley, S., Bradon, P., Evans, J., & Newstead, S. (2002). Approaches to modeling item-generative tests. Approaches to Modeling Item Generative Tests. In *Item Generation for Test Development* (pp. 85-104). Routledge.

Dorn, S. (2010). The political dilemmas of formative assessment. *Exceptional Children*, *76*(3), 325-337.

Downing, S. M., & Haladyna, T. M. (Eds.). (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.

Enright, M. K., & Sheehan, K. M. (2002). Modeling the difficulty of quantitative reasoning items: Implications for item generation. *Item generation for test development*, 129-157.

Epstein, J. I. (1997). The effects of different types of feedback on learning verbal reasoning in computer based instruction (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 9727464)

Fossati, D., Di Eugenio, B., Ohlsson, S., Brown, C., & Chen, L. (2015). Data driven automatic feedback generation in the iList intelligent tutoring system. *Technology, Instruction, Cognition and Learning*, *10*(1), 5-26.

Furnham, A., Batey, M., & Martin, N. (2011). How would you like to be evaluated? The correlates of students' preferences for assessment methods. *Personality and Individual Differences*, *50*(2), 259-263.

Gambhir, M., Broad, K., Evans, M., & Gaskell, J. (2008). Characterizing initial teacher education in Canada: Themes and issues. *Toronto: Ontario Institute for Studies in Education*.

Gappa, J. M., Austin, A. E., & Trice, A. G. (2007). *Rethinking faculty work: Higher education's strategic imperative*. Jossey-Bass.

Gates, D. (2008). Generating look-back strategy questions from expository texts. In *The Workshop on the Question Generation Shared Task and Evaluation Challenge, NSF, Arlington, VA.*

Geng, G. (2005). Data Structure—C Language description. *Xi'an Electronic Science and Technology University Press, Xi'an*.

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning and Assessment*, *7*(2).

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice*, *32*(3), 36-50.

Gierl, M. J., Ball, M. M., Vele, V., & Lai, H. (2015, June). A Method for Generating Nonverbal Reasoning Items Using n-Layer Modeling. In *International Computer Assisted Assessment Conference* (pp. 12-21). Springer, Cham.

Gierl, M. J. & Lai, H. (2016a). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of test development* (2nd edition, pp. 410-429). New York: Routledge.

Gierl, M. J., & Lai, H. (2016b). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice*, *35*(4), 6-20.

Gierl, M. J., & Lai, H. (2016c). The role of cognitive models in automatic item generation. *The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications*, 124.

Gierl, M. J., & Lai, H. (2018). Using Automatic Item Generation to Create Solutions and Rationales for Computerized Formative Testing. *Applied Psychological Measurement*, *42*(1), 42–57.

Gierl, M., Bulut, O., & Zhang, X. (2018). Using Computerized Formative Testing to Support Personalized Learning in Higher Education: An Application of Two Assessment Technologies. In *Digital Technologies and Instructional Design for Personalized Learning* (pp. 99-119). IGI Global.

Goeters, K. M. & Lorenz, B. (2002). On the implementation of item generation principles in the design of aptitude testing in aviation. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development*. Mahwah, NJ: Erlbaum.

Gordijn, J., & Nijhof, W. J. (2002). Effects of complex feedback on computer-assisted modular instruction. *Computers and Education*, *39*, 183–200.

Graf, E. A., & Fife, J. H. (2013). Difficulty modeling and automatic generation of quantitative items: Recent advances and possible next steps. *Automatic Item Generation: Theory and Practice*, 157-179.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hall, K. A., Adams, M., & Tardibuono, J. (1968). Gradient-and full-response feedback in computer-assisted instruction. *The Journal of Educational Research*, *61*(5), 195-199.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of educational research*, *77*(1), 81-112.

Heilman, M., & Smith, N. A. (2009). *Question generation via overgenerating transformations and ranking* (No. CMU-LTI-09-013). Carnegie-Mellon Univ Pittsburgh Pa Language Technologies Inst.

Heilman, M. (2011). *Automatic factual question generation from text* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database.

Herman, J. L., Osmundson, E., Ayala, C., Schneider, S., & Timms, M. (2006). The Nature and Impact of Teachers' Formative Assessment Practices. CSE Technical Report 703. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*.

Hornke, L. (2002). Item generation models for higher cognitive functions. In S. Irvine & P. Kyllonen (Eds.), *Item generation for test development* (p.159-178). Mahwah, NJ: Erlbaum.

Hoska, D. M. (1993). Motivating learners through CBI feedback: Developing a positive learner perspective. *Interactive instruction and feedback*, 105-132.

Hwang, G. J., & Chang, H. F. (2011). A formative assessment-based mobile learning approach to improving the learning attitudes and achievements of students. *Computers & Education*, *56*(4), 1023-1031.

Ifenthaler, D. (2010). Bridging the gap between expert-novice differences: The model-based feedback approach. *Journal of Research on Technology in Education*, *43*, 103–117.

Indurkhya, N., & Damerau, F. J. (Eds.). (2010). *Handbook of natural language processing* (Vol. 2). CRC Press.

LaDuca, A., Staples, W. I., Templeton, B., & Holzman, G. B. (1986). Item modelling procedure for constructing content-equivalent multiple-choice questions. *Medical education*, *20*(1), 53-56.

Lane, S., Raymond, M., Haladyna, R., & Downing, S. (2016). Test development process. In S. Lane, M.Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 3-

18). New York, NY: Routledge.

Serban, I. V., García-Durán, A., Gulcehre, C., Ahn, S., Chandar, S., Courville, A., & Bengio, Y. (2016). Generating factoid questions with recurrent neural networks: The 30m factoid question-answer corpus. *arXiv preprint arXiv:1603.06807*.

Jacquemin, C., & Tzoukermann, E. (1999). NLP for term variant extraction: synergy between morphology, lexicon, and syntax. In *Natural language information retrieval* (pp. 25-74). Springer, Dordrecht.

Kane, M. (2016). Validation Strategies. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 64-80). New York, NY: Routledge.

Kim, J. Y. L., & Phillips, T. L. (1991). The effectiveness of two forms of corrective feedback in diabetes education. *Journal of Computer-Based Instruction*, *18*(1), 14–18.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational measurement: Issues and practice*, *30*(4), 28-37.

Klein, D., & Manning, C. D. (2003). Fast exact inference with a factored model for natural language parsing. In *Advances in neural information processing systems* (pp. 3-10).

Kopp, V., Stark, R., & Fischer, M. R. (2008). Fostering diagnostic knowledge through computer-supported, case-based worked examples: Effects of erroneous examples and feedback. *Medical Education*, *42*, 823–829.

Kramarski, B., & Zeichner, O. (2001). Using technology to enhance mathematical reasoning: Effects of feedback and self-regulation learning. *Educational Media International*, *38*, 77–82.

Kulhavy, R. W., White, M. T., Topp, B. W., Chan, A. L., & Adams, J. (1985). Feedback complexity and corrective efficiency. *Contemporary educational psychology*, *10*(3), 285-291.

Lai, H. (2013). *Developing a Framework and Demonstrating a Systematic Process for Generating Medical Test Items* (Unpublished doctoral dissertation). University of Alberta, Edmonton.

Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research and Development*, *58*, 629–648.

Lin, H. (2006). *The effect of questions and feedback used to complement static and animated visualization on tests measuring different educational objectives* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3318901)

Lindberg, D., Popowich, F., Nesbit, J., & Winne, P. (2013). Generating natural language questions to support learning on-line. In *Proceedings of the 14th European Workshop on Natural Language Generation* (pp. 105-114).

Lipnevich, A. A., & Smith, J. K. (2009). Effects of differential feedback on students' examination performance. *Journal of Experimental Psychology: Applied*, *15*, 319–333.

Mannem, P., Prasad, R., & Joshi, A. (2010, June). Question generation from paragraphs at UPenn: QGSTEC system description. In *Proceedings of QG2010: The Third Workshop on Question Generation* (pp. 84-91).

Mason, B. J., & Bruning, R. (2001). *Providing feedback in computer-based instruction: What the research tells us*. CLASS Research Report No.9. Center for Instructional Innovation, University of Nebraska–Lincoln.

Mazingo, D. E. (2006). *Identifying the relationship between feedback provided in computer-assisted instructional modules, science self-efficacy, and academic achievement* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3263234)

Merril, J. (1987). Levels of questioning and forms of feedback: Instructional factors in courseware design. *Journal of Computer-Based Instruction*, *14*(1), 18–22.

Miller, T. (2009). Formative computer-based assessment in higher education: The effectiveness of feedback in supporting student learning. *Assessment & Evaluation in Higher Education*, *34*(2), 181-192.

Mitkov, R., & Ha, L. A. (2003). Computer-aided generation of multiple-choice tests. In *Proceedings of the HLT-NAACL 03 workshop on Building educational applications using natural language processing-Volume 2* (pp. 17-22). Association for Computational Linguistics.

Moreno, N. (2007). *The effects of type of task and type of feedback on L2 development in call* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3302088)

Moreno, R. (2004). Decreasing cognitive load for novice students: Effects of explanatory versus corrective feedback in discovery-based multimedia. *Instructional science*, *32*(1-2), 99-113.

Moreno, R., & Valdez, A. (2005). Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development*, *53*, 35–45.

Morrison, G. R., Ross, S. M., Gopalakrishnan, M., & Casey, J. (1995). The effects of feedback and incentives on achievement in computer-based instruction. *Contemporary Educational Psychology*.

Moreno, R., Martínez, R. J., & Muñiz, J. (2006). New guidelines for developing multiple-choice items. *Methodology*, *2*(2), 65-72.

Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema*, *27*(4), 388-394.

Mostow, J., & Chen, W. (2009). Generating instruction automatically for the reading strategy of self-questioning. In *Proceedings of the 2009 Conference on Artificial Intelligence in Education: Building Learning Systems That Care: From Knowledge Representation to Affective Modelling*. pages 465–472.

Munyofu, M. (2008). *Effects of varied enhancement strategies (chunking, feedback, gaming) in complementing animated instruction in facilitating different types of learning objectives* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3414357)

Murphy, P. (2007). Reading comprehension exercises online: The effects of feedback, proficiency and interaction. *Language Learning and Technology*, 11, 107–129.

Murphy, P. (2010). Web-based collaborative reading exercises for learners in remote locations: The effects of computer-mediated feedback and interaction via computer mediated communication. *ReCALL*, *22*, 112–134.

Nagata, N. (1993). Intelligent computer feedback for second language instruction. *Modern Language Journal*, *77*, 330–339.

Nagata, N., & Swisher, M. V. (1995). A study of consciousness-raising by computer: The effect of metalinguistic feedback on second language learning. *Foreign Language Annals*, *28*, 337–347.

Narciss, S., & Huth, K. (2004). How to design informative tutoring feedback for multi- media learning. In H. M. Niegemann, D. Leutner, & R. Brunken (Ed.), *Instructional design for multimedia learning* (pp. 181–195). Munster, NY: Waxmann.

Neri, A., Cucchiarini, C., & Strik, H. (2008). The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch. *ReCALL*, *20*(2), 225-243.

Papa, F. J., Aldrich, D. A. V. I. D., & Schumacker, R. E. (1999). The effects of immediate online feedback upon diagnostic performance. *Academic Medicine*, *74*(10), S16-8.

Peat, M., & Franklin, S. (2002). Supporting student learning: the use of computer-based formative assessment modules. *British Journal of Educational Technology*, *33*(5), 515–523. https://doi.org/10.1111/1467-8535.00288

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J. H., & Jurafsky, D. (2005). Support vector learning for semantic argument classification. *Machine Learning*, *60*(1-3), 11-39.

Pridemore, D. R., & Klein, J. D. (1991). Control of feedback in computer-assisted instruction. *Educational Technology Research and Development*, *39*, 27–32.

Pridemore, D. R., & Klein, J. D. (1995). Control of practice and level of feedback in computer-based instruction. *Contemporary Educational Psychology*, *20*, 444–450.

Quinton, S., & Smallbone, T. (2010). Feeding forward: using feedback to promote student reflection and learning–a teaching model. *Innovations in Education and Teaching International*, *47*(1), 125-135.

Rodriguez, M. (2016). Selected-Response Item Development. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 259-273). New York, NY: Routledge.

Rosa, E. M., & Leow, R. P. (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. *The Modern Language Journal*, *88*(2), 192-216.

Roos, L. L., Wise, S. L., & Plake, B. S. (1997). The role of item feedback in self-adapted testing. *Educational and Psychological Measurement*, *57*(1), 85-98.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of research in science teaching*, *44*(1), 57-84.

Rus, V., Graesser, A., & Cai, Z. (2008). Question Generation: Example of A Multi-year Evaluation Campaign. *Workshop on the Question Generation Shared Task and Evaluation Challenge*, (January). Retrieved from http://www.cs.memphis.edu/~vrus/questiongeneration/5-RusEtAl-QG08.pdf

Bhatia, A. S., Kirti, M., & Saha, S. K. (2013). Automatic generation of multiple choice questions using wikipedia. In *International Conference on Pattern Recognition and Machine Intelligence* (pp. 733-738). Springer, Berlin, Heidelberg.

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. *Language Learning*, *54*(1), 35-78.

Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, *78*(1), 153-189.

Susanti, Y., Iida, R., & Tokunaga, T. (2015, May). Automatic generation of english vocabulary tests. In *CSEDU (1)* (pp. 77-87).

U.S. Department of Education, National Center for Education Statistics. (2018). *Digest of Education Statistics, 2016*, Chapter 3.Retrieved from https://nces.ed.gov/programs/digest/d16/ch_3.asp

Valdez, A. J. (2009). *Encouraging mindful feedback processing: Computer-based instruction in descriptive statistics* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3329482)

VanEvera, W. C. (2003). *Achievement and motivation in the middle school science classroom: The effects of formative assessment feedback*. (Doctoral dissertation. Dissertations & theses: Full textdatabase (Publication No. AAT 3110082).

van der Kleij, F. M., Timmers, C. F., & Eggen, T. J. (2011). The effectiveness of methods for providing written feedback through a computer-based assessment for learning: A systematic review. *Cadmo*.

van der Kleij, F. M., Eggen, T. J., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education*, *58*(1), 263-272.

Van der Kleij, F. M., Feskens, R. C., & Eggen, T. J. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of educational research*, *85*(4), 475-511.

von Davier, M. (2018). Automated Item Generation with Recurrent Neural Networks. *Psychometrika*, 1-11.

Williams, S. E. (1997, March). Teachers' written comments and students' responses: A socially constructed interaction. Proceedings of the annual meeting of the Conference on College Composition and Communication, Phoenix, AZ.

Wilson, K., Boyd, C., Chen, L., & Jamal, S. (2011). Improving student performance in a first-year geography course: Examining the importance of computer-assisted formative assessment. *Computers & Education*, *57*(2), 1493-1500.

Wise, L., & Plake, B. (2016). Test Design and Development Following the Standards for Educational and Psychological Testing. In S. Lane, M. Raymond, & T. Haladyna (Eds.), Handbook of test development (2nd ed., pp. 19-39). New York, NY: Routledge.

Xu, M. (2009). *An investigation of the effectiveness of intelligent elaborative feed- back afforded by pedagogical agents on improving young Chinese language learners' vocabulary acquisition* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (UMI No. 3359045)

Yao, X., & Zhang, Y. (2010, June). Question generation with minimal recursion semantics. In *Proceedings of QG2010: The Third Workshop on Question Generation* (pp. 68-75).

Yao, X., Bouma, G., & Zhang, Y. (2012). Semantics-based question generation and implementation. *Dialogue & Discourse*, *3*(2), 11-42.

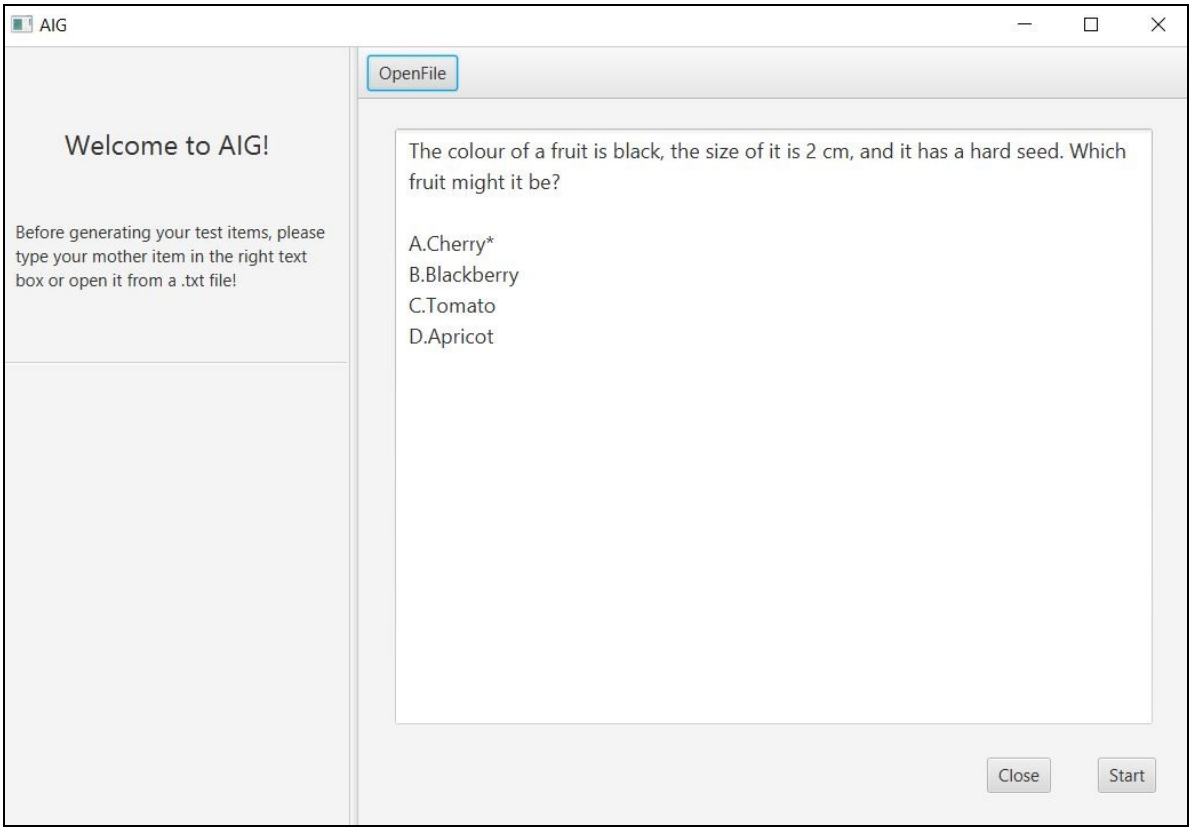
Yuan, X., Wang, T., Gulcehre, C., Sordoni, A., Bachman, P., Subramanian, S., Trischler, A. (2017). Machine Comprehension by Text-to-Text Neural Question Generation. Retrieved from https://arxiv.org/pdf/1705.02012.pdf

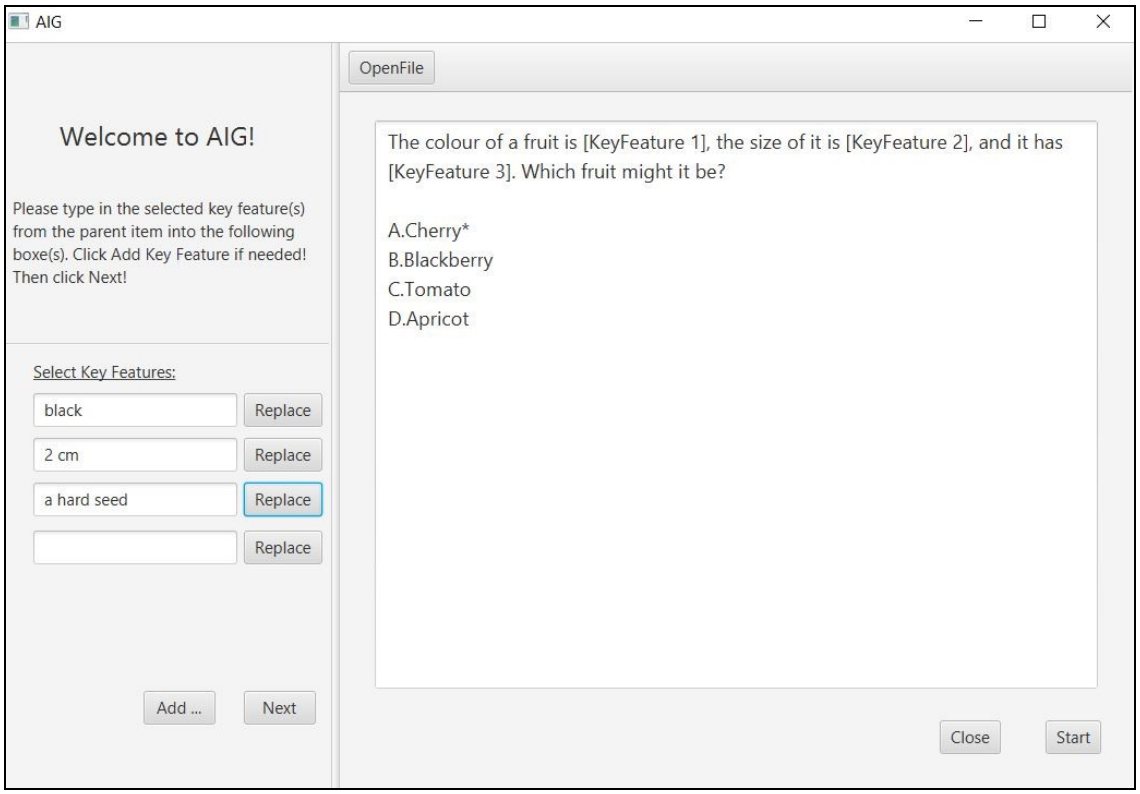
Zhang, X., & Gierl, M. (2016). A Model-Based Method for Content Validation of Automatically Generated Test Items. *Journal of Educational Issues*, *2*(2), 184-202.

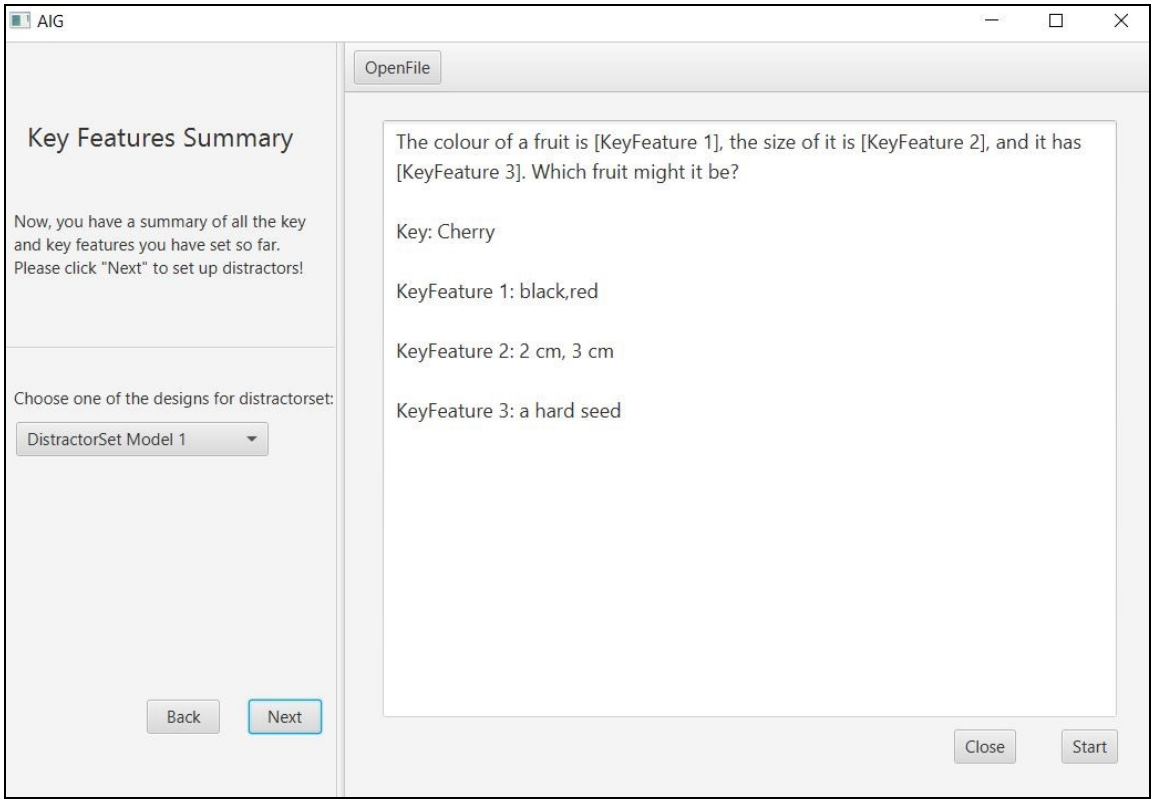
Zhang, X., Gierl, M. (2018). Automatic Item Generation in creating content for educational assessments: A comprehensive review. *Computers &Education* (under review).

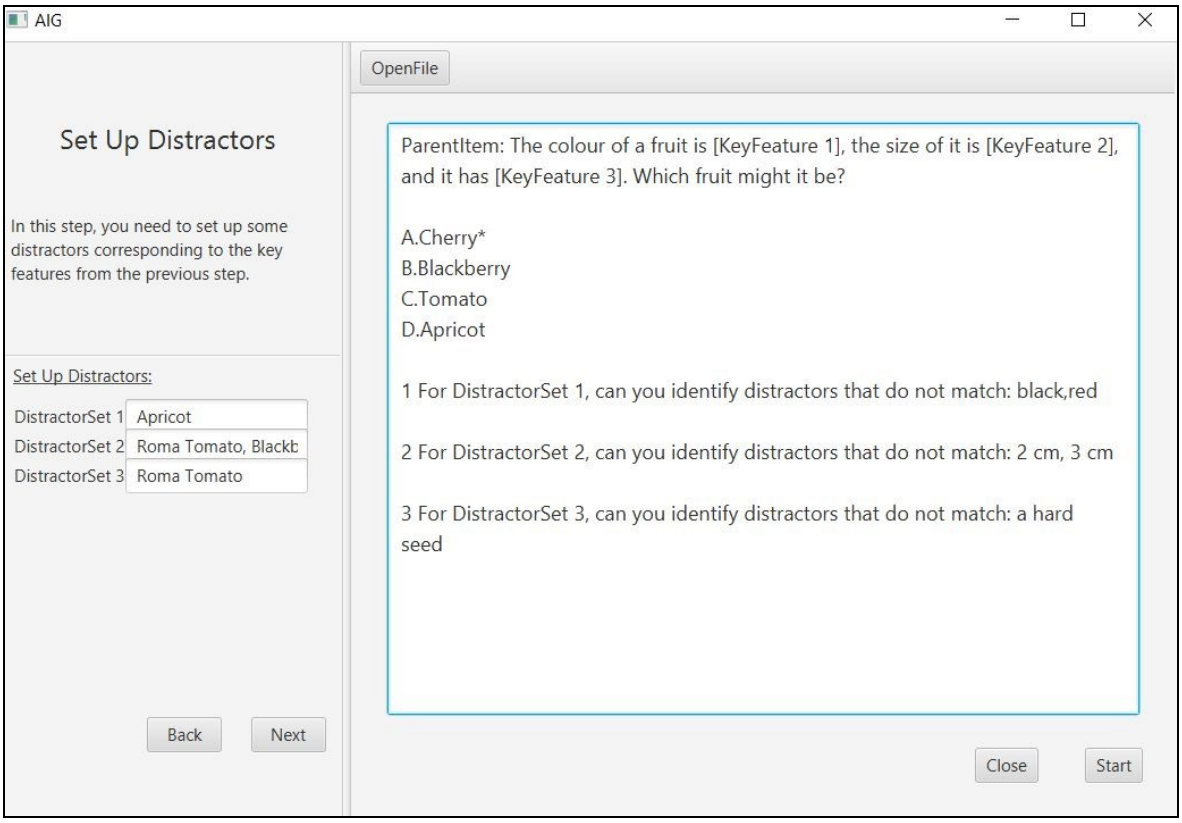
Appendix A

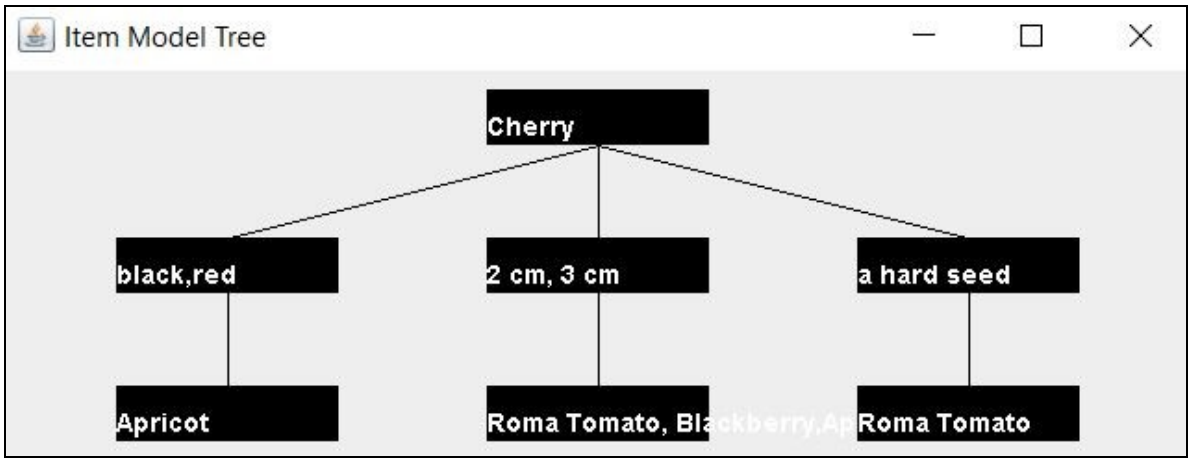
Demonstration with an Example (selected screen shots of using HIM_AIG)

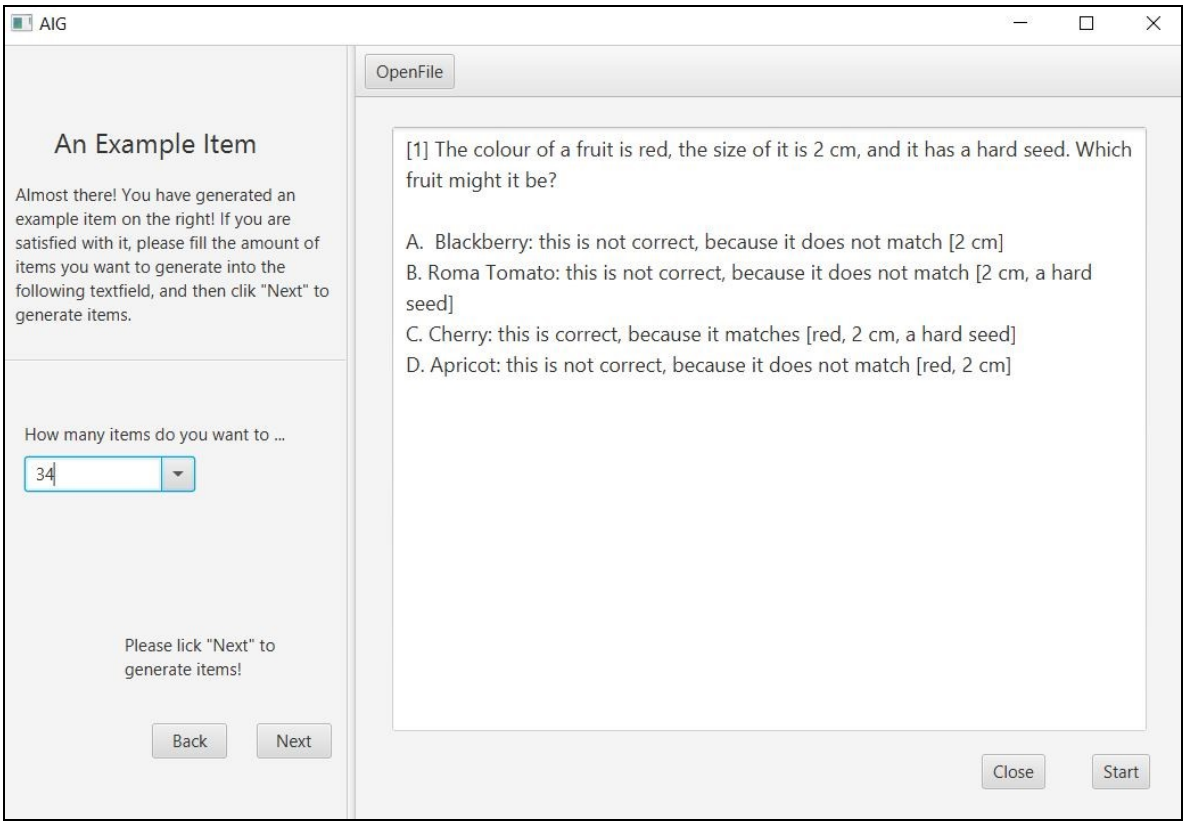


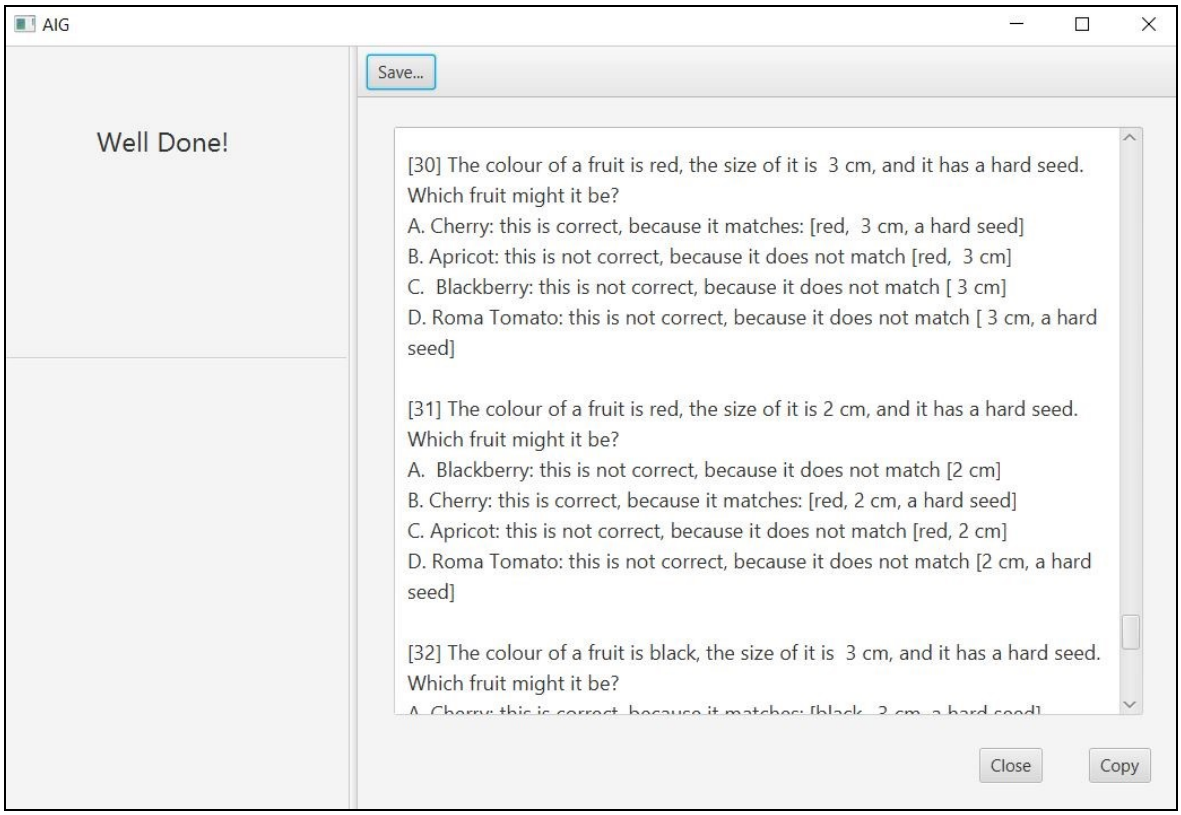

AIG
OpenFile
Welcome to AIG!
Please type in the selected key feature(s) from the parent item into the following boxe(s). Click Add Key Feature if needed! Then click Next!
Select Key Features:
black
Replace
2 cm
Replace
a hard seed
Replace
Replace
Add ...
Next
The colour of a fruit is [KeyFeature 1], the size of it is [KeyFeature 2], and it has [KeyFeature 3]. Which fruit might it be?
A.Cherry*
B.Blackberry
C.Tomato
D.Apricot
Close
Start


AIG
OpenFile
Key Features Summary
Now, you have a summary of all the key and key features you have set so far. Please click "Next" to set up distractors!
Choose one of the designs for distractorset:
DistractorSet Model 1
Back
Next
The colour of a fruit is [KeyFeature 1], the size of it is [KeyFeature 2], and it has [KeyFeature 3]. Which fruit might it be?
Key: Cherry
KeyFeature 1: black,red
KeyFeature 2: 2 cm, 3 cm
KeyFeature 3: a hard seed
Close
Start


AIG
OpenFile
Set Up Distractors
In this step, you need to set up some distractors corresponding to the key features from the previous step.
Set Up Distractors:
DistractorSet 1
Apricot
DistractorSet 2
Roma Tomato, Blackb
DistractorSet 3
Roma Tomato
Back
Next
ParentItem: The colour of a fruit is [KeyFeature 1], the size of it is [KeyFeature 2], and it has [KeyFeature 3]. Which fruit might it be?
A.Cherry*
B.Blackberry
C.Tomato
D.Apricot
1 For DistractorSet 1, can you identify distractors that do not match: black,red
2 For DistractorSet 2, can you identify distractors that do not match: 2 cm, 3 cm
3 For DistractorSet 3, can you identify distractors that do not match: a hard seed
Close
Start

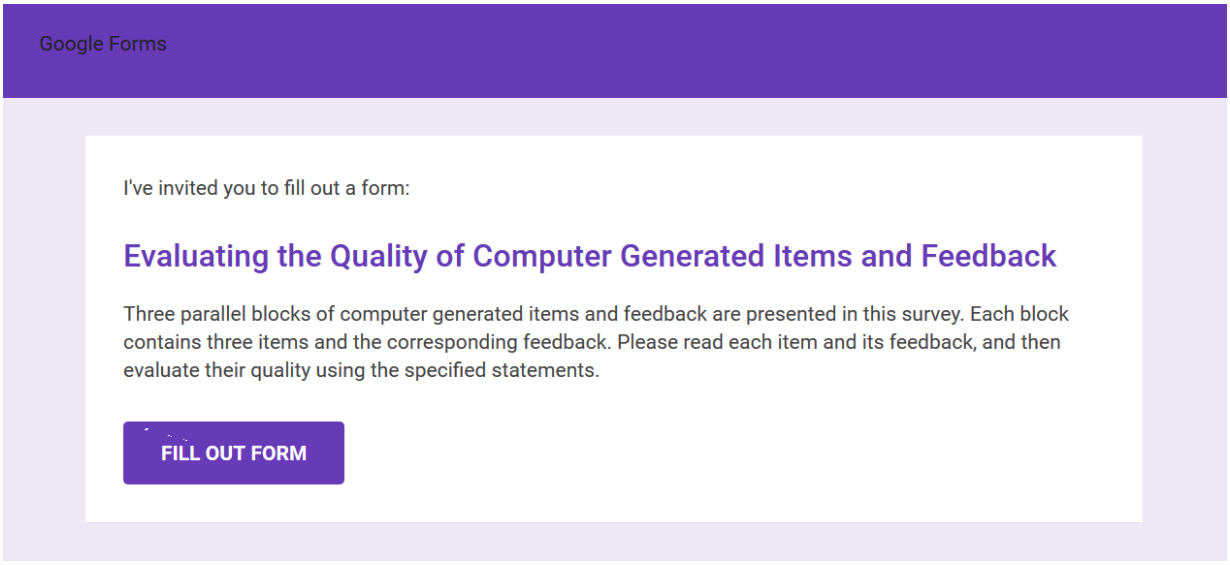

Item Model Tree
Cherry
black,red
2 cm, 3 cm
a hard seed
Apricot
Roma Tomato, Blackberry,Ap
Roma Tomato

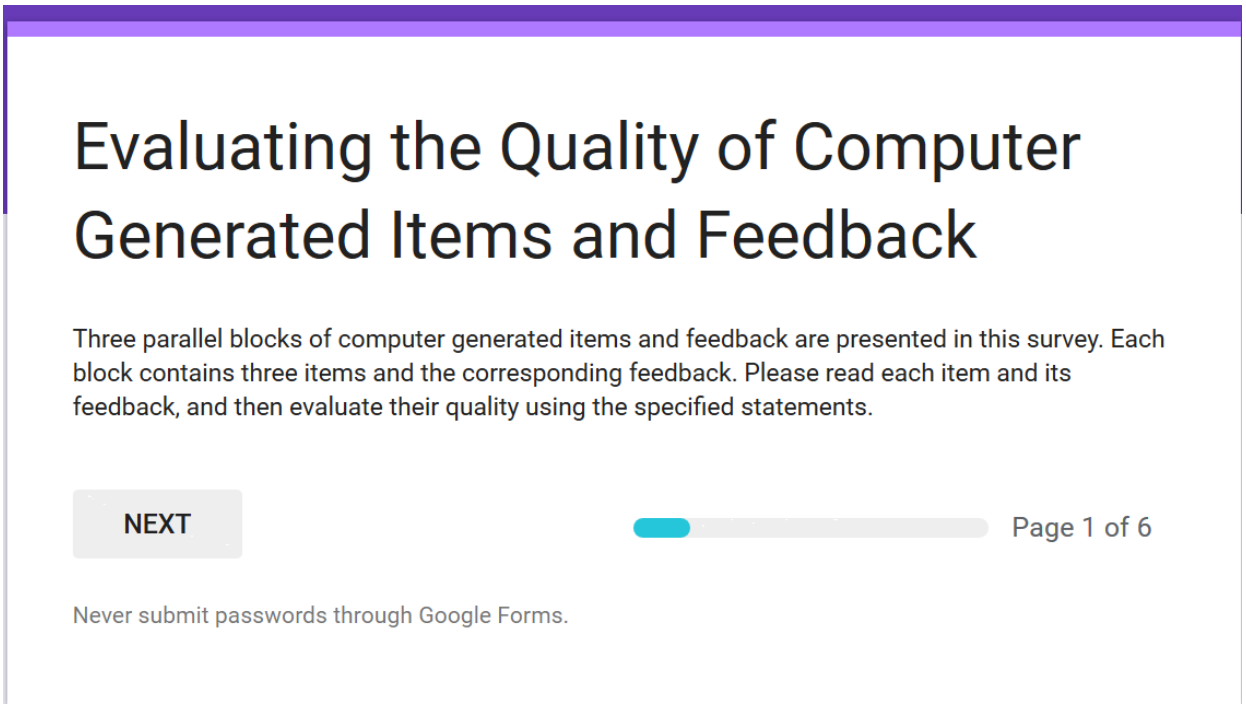

AIG
OpenFile
An Example Item
Almost there! You have generated an example item on the right! If you are satisfied with it, please fill the amount of items you want to generate into the following textfield, and then clik "Next" to generate items.
How many items do you want to ...
34
Please lick "Next" to generate items!
Back
Next
[1] The colour of a fruit is red, the size of it is 2 cm, and it has a hard seed. Which fruit might it be?
A. Blackberry: this is not correct, because it does not match [2 cm]
B. Roma Tomato: this is not correct, because it does not match [2 cm, a hard seed]
C. Cherry: this is correct, because it matches [red, 2 cm, a hard seed]
D. Apricot: this is not correct, because it does not match [red, 2 cm]
Close
Start


AIG
Save...
Well Done!
[30] The colour of a fruit is red, the size of it is 3 cm, and it has a hard seed. Which fruit might it be?
A. Cherry: this is correct, because it matches: [red, 3 cm, a hard seed]
B. Apricot: this is not correct, because it does not match [red, 3 cm]
C. Blackberry: this is not correct, because it does not match [3 cm]
D. Roma Tomato: this is not correct, because it does not match [3 cm, a hard seed]
[31] The colour of a fruit is red, the size of it is 2 cm, and it has a hard seed. Which fruit might it be?
A. Blackberry: this is not correct, because it does not match [2 cm]
B. Cherry: this is correct, because it matches: [red, 2 cm, a hard seed]
C. Apricot: this is not correct, because it does not match [red, 2 cm]
D. Roma Tomato: this is not correct, because it does not match [2 cm, a hard seed]
[32] The colour of a fruit is black, the size of it is 3 cm, and it has a hard seed. Which fruit might it be?
Close
Copy

Appendix B

Survey





Evaluating the Quality of Computer Generated Items and Feedback

Block 1

1. Please read item 1 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 1
Example 1: "Distinguish between relevant and irrelevant numbers in a math word problem" Example 2: "Write a research paper" Example 3: "Check if conclusions follow the observed data" Which of the following trios of Bloom's levels best describes examples 1, 2 and 3, respectively? A*. Analyze; Create; Evaluate B. Analyze; Evaluate; Analyze C. Evaluate; Create; Evaluate D. Evaluate; Evaluate; Analyze
Feedback
• A is correct, because it matches every example: "Distinguish between relevant and irrelevant numbers in a math word problem", "Write a research paper", "Check if conclusions follow the observed data" • B is incorrect, because it does not match the following example(s): "Write a research paper", "Check if conclusions follow the observed data" • C is incorrect, because it does not match the following example(s): "Distinguish between relevant and irrelevant numbers in a math word problem" • D is incorrect, because it does not match the following example(s): "Distinguish between relevant and irrelevant numbers in a math word problem", "Write a research paper", "Check if conclusions follow the observed data"

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

2. Please read item 2 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 2
Which of the following pairs best describe the levels at which "difficulty" and "reliability" are analyzed, respectively? A. test level; test level B. *item level; test level C. test level; item level D. item level; item level
Feedback
• A is incorrect, because it does not match the following term(s): "difficulty" • B is correct, because it matches every term: "difficulty", "reliability" • C is incorrect, because it does not match the following term(s): "difficulty", "reliability" • D is incorrect, because it does not match the following term(s): "reliability"

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

3. Please read item 3 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 3
Which tool has the medium objectivity in scoring, medium authenticity and complexity, and medium time requirement for developing? A. *Constructed-response items B. Selected-response items C. Performance assessment tasks D. Portfolio assessment
Feedback
• A is correct, because it matches every feature: medium objectivity in scoring, medium authenticity and complexity, medium time requirement for developing • B is incorrect, because it does not match the following feature(s): medium objectivity in scoring, medium authenticity and complexity • C is incorrect, because it does not match the following feature(s): medium authenticity and complexity, medium time requirement for developing • D is incorrect, because it does not match the following feature(s): medium authenticity and complexity, medium time requirement for developing

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

BACK NEXT

Evaluating the Quality of Computer Generated Items and Feedback

Block 2

1. Please read item 4 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 4

Example 1: "Judge which of two methods is the best way to solve a problem "
Example 2: "Structure historical evidence for and against a particular historical explanation"
Example 3: "Generate a hypothesis to account for observed phenomenon"
Which of the following trios of Bloom's levels best describes examples 1, 2 and 3, respectively?

A. *Evaluate; Analyze; Create
B. Analyze; Evaluate; Create
C. Evaluate; Analyze; Analyze
D. Analyze; Evaluate; Analyze

Feedback

- A is correct, because it matches every example: "Judge which of two methods is the best way to solve a problem ", "Structure historical evidence for and against a particular historical explanation", "Generate a hypothesis to account for observed phenomenon"
- B is incorrect, because it does not match the following example(s): "Judge which of two methods is the best way to solve a problem ", "Structure historical evidence for and against a particular historical explanation"
- C is incorrect, because it does not match the following example(s): "Generate a hypothesis to account for observed phenomenon"
- D is incorrect, because it does not match the following example(s): "Judge which of two methods is the best way to solve a problem ", "Structure historical evidence for and against a particular historical explanation", "Generate a hypothesis to account for observed phenomenon"

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

2. Please read item 5 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 5

Which of the following pairs best describe the levels at which "descriptive analysis of test scores" and "discrimination" are analyzed, respectively?

A. test level; test level
B. item level; test level
C. *test level; item level
D. item level; item level

Feedback

- A is incorrect, because it does not match the following term(s): "discrimination"
- B is incorrect, because it does not match the following term(s): "descriptive analysis of test scores", "discrimination"
- C is correct, as it matches every term: "descriptive analysis of test scores", "discrimination"
- D is incorrect, because it does not match the following term(s): "descriptive analysis of test scores"

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

3. Please read item 6 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 6

Which tool has the high objectivity in scoring, low authenticity and complexity, and low time requirement for developing?

A. Performance assessment tasks
B. Portfolio assessment
C. *Selected-response items
D. Constructed-response items

Feedback

- A is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, low time requirement for developing
- B is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, low time requirement for developing
- C is correct, as it matches every feature: high objectivity in scoring, low authenticity and complexity, and low time requirement for developing
- D is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, low time requirement for developing

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

BACK NEXT

Evaluating the Quality of Computer Generated Items and Feedback

Block 3

1. Please read item 7 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 7
Example 1: "Distinguish between relevant and irrelevant numbers in a math word problem" Example 2: "Build habitats for a specific purpose" Example 3: "Check if conclusions follow the observed data" Which of the following trios of Bloom's levels best describes examples 1, 2 and 3, respectively? A. Evaluate; Create; Evaluate B. *Analyze; Create; Evaluate C. Evaluate; Evaluate; Analyze D. Analyze; Evaluate; Analyze
Feedback
• A is incorrect, because it does not match the following example(s): "Distinguish between relevant and irrelevant numbers in a math word problem" • B is correct, as it matches every example: "Distinguish between relevant and irrelevant numbers in a math word problem", "Build habitats for a specific purpose", "Check if conclusions follow the observed data" • C is incorrect, because it does not match the following example(s): "Distinguish between relevant and irrelevant numbers in a math word problem", "Build habitats for a specific purpose", "Check if conclusions follow the observed data" • D is incorrect, because it does not match the following example(s): "Build habitats for a specific purpose", "Check if conclusions follow the observed data"

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

2. Please read item 8 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

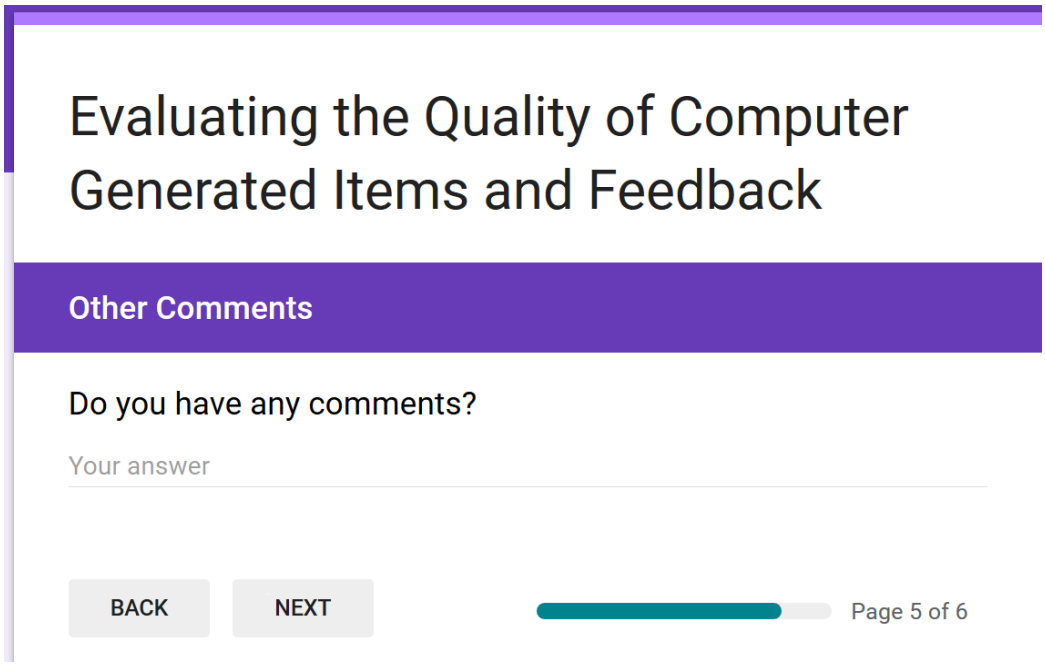
Item 8
Which tool has the high objectivity in scoring, low authenticity and complexity, and medium time requirement for developing? A. Portfolio assessment B. *Selected-response items C. Performance assessment tasks D. Constructed-response items
Feedback
• A is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, medium time requirement for developing • B is correct, as it matches every feature: high objectivity in scoring, low authenticity and complexity, medium time requirement for developing • C is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity, medium time requirement for developing • D is incorrect, because it does not match the following feature(s): high objectivity in scoring, low authenticity and complexity

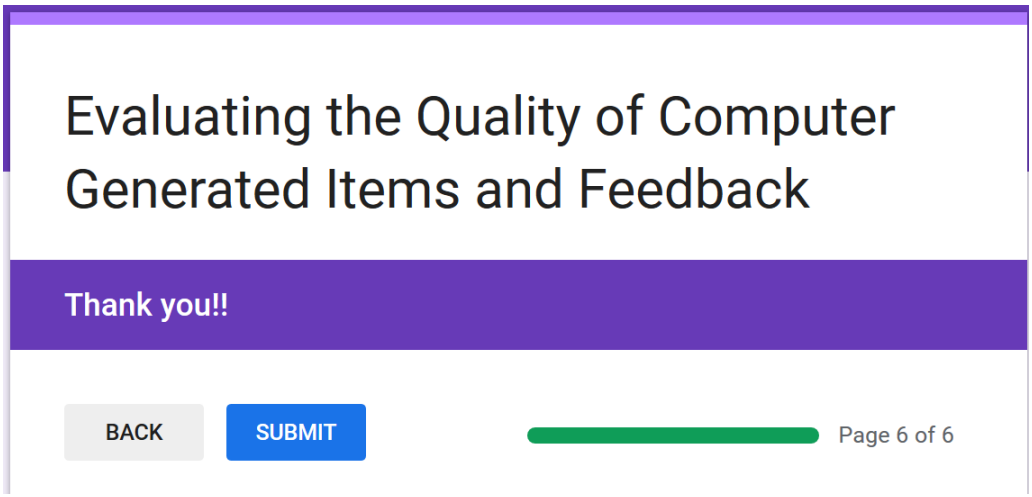
	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○

3. Please read item 9 and its feedback, and indicate the extent to which you agree or disagree with each of the five statements that follow.

Item 9
Which of the following pairs best describe the levels at which "descriptive analysis of test scores" and "distractor functioning" are analyzed, respectively? A. test level; test level B. item level; item level C. item level; test level D. *test level; item level
Feedback
• A is incorrect, because it does not match the following term(s): "distractor functioning" • B is incorrect, because it does not match the following term(s): "descriptive analysis of test scores" • C is incorrect, because it does not match the following term(s): "descriptive analysis of test scores", "distractor functioning" • D is correct, because it matches every term: "descriptive analysis of test scores", "distractor functioning"

	strongly disagree	disagree	agree	strongly agree
The item presents the main idea/intended content clearly	○	○	○	○
The item provides plausible options	○	○	○	○
The item provides independent options which give no clues to the correct response	○	○	○	○
The feedback provides an elaborated explanation of the item	○	○	○	○


Evaluating the Quality of Computer Generated Items and Feedback
Other Comments
Do you have any comments?
Your answer
BACK
NEXT
Page 5 of 6


Evaluating the Quality of Computer Generated Items and Feedback
Thank you!!
BACK
SUBMIT
Page 6 of 6