Macrothink Institute™

Journal of Educational Issues
ISSN 2377-2263
2018, Vol. 4, No. 1

Evaluation of Automatic Item Generation Utilities in Formative Assessment Application for Korean High School Students

Jaehwa Choi (Corresponding author)

Assessment, Testing, and Measurement, The George Washington University

2134 G St NW, Washington, DC 20052, USA

Tel: 1-202-994-2602 E-mail: jaechoi@gwu.edu

HeeKyoung Kim

Division of Educational Evaluation, Korea Institute for Curriculum and Evaluation

8, Gyohak-ro, Deoksan-myeon, Jincheon-gun, Chungcheongbuk-do, 27873, Korea

Tel: 82-43-931-0224 E-mail: heekyoung@kice.re.kr

Seohong Pak

Psychometrics and Data Analysis, National Board of Medical Examiners

3750 Market Street, Philadelphia, PA 19104, USA

Tel: 1-215-590-9857 E-mail: spak@nbme.org

Received: February 9, 2018 Accepted: March 16, 2018 Published: March 19, 2018

doi:10.5296/jei.v4i1.12630 URL: https://doi.org/10.5296/jei.v4i1.12630

Abstract

The recent interests in research in the assessment field have been rapidly shifting from decision-maker-centered assessments to learner-centered assessments (*i.e.*, diagnostic and/or formative assessments). In particular, it is a very important research topic in this field to analyze how these learner-centered assessments are developed more practical and valid by combining information (or intelligent) and communication technologies (ICT) and

psychometric advances, and how these systems contribute to learning. Automatic Item Generation (AIG), which is in the spotlight recently, is a representative example of building a next-generation assessment theory and practice by integrating ICT into psychometrics.

AIG has very broad and promising features in mass production, intelligent item calibration and management, especially learner-centered assessment, and so on. However, these claims have not yet been fully validated in real education settings, and it is in dire need of evaluating the utilities of AIG applications in practical educational settings. Based on these needs, the purposes of this research are, firstly, to introduce the benefits and challenges of AIG to properly assess AIG's utilities. And, secondly, this study provides empirical evaluation results of AIG's utilities within a formative assessment system for Korean high school students and teachers.

The results of this study have important implications in that it is the first empirical study to evaluate the usefulness of AIG in the formation evaluation application of the actual educational environment, Korean high school setting. The results of the theoretical and empirical evaluation of the AIG will also be useful to researchers and practitioners who wish to evaluate the benefits and issues of future AIG-based educational services or theory development.

Keywords: Automatic item generation, Formative assessment, Measurement, Testing, Evaluation, Computer-based testing, Computer adaptive testing

1. Introduction

A recent paradigm in assessment focuses more on helping students understand where they currently are and what is necessary to improve their learning. Therefore, it becomes important for assessment to identify the bottleneck of student learning and provide feedback than to rank students or trig competition among students. That is, the assessment *for* learning or *as* learning is preferred to the assessment for *evaluating*. In order to meet the needs of current paradigm in assessment, a formative assessment integrated into high school homework system has been discussed as a great candidate.

Formative assessment systems are well-designed and well-used and can provide a variety of benefits to both teachers and students. For example, formative assessments implemented at the end of every unit can help teachers find appropriate learning methods that help them to recognize the student's achievement status and improve student achievement. Also, the most valuable gain from formative assessments could be helping students understand their own learning, and develop appropriate strategies for "learning to learn" (OECD & CERI, 2008). Despite these advantages of formative assessment, it is impractical to expect all teachers in the current school environment to develop qualified assessments, analyze results quickly, and provide feedback to students on time. It is urgent to develop and support information and communication technology (ICT) based system for modeling evaluation so that teachers can use the evaluation items whenever necessary.

There are a few considerations when developing such a formative system. First, the system must have an item bank containing a sufficient number of qualified items. Second, these items

not only assess students' learning status, but they can also provide practice items repeatedly as needed. Third, students should get the results (*i.e.*, what is right and wrong) and the system should provide detailed feedback on students' strengths and weaknesses.

Automatic Item Generation (AIG; Choi, 2017; Embretson, 1998; Gierl & Haladyna, 2012; Irvine & Kyllonen, 2002) can be a promising solution to meet all three of the prerequisites mentioned above. With AIG technique, item model (or item template) developed by experts (*i.e.*, item model writers) is used in producing clone items (also called isomorphs) via a computer-based application. AIG is able to produce a large amount of new items and ensures that there are enough items in the item bank for formative assessment. Another reason that AIG is a good solution for a formative assessment system is that it allows one to have as many qualified items as possible, regardless of the psychometric measurement model used in the scoring method. The system geared with AIG can also increase the accuracy of student assessment with a sufficient amount of items.

Another benefit of AIG is related to immediate feedback. In other words, incorrect items can be duplicated as well as descriptions, ensuring that students fully understand the essential elements included in the parent items. By repeatedly providing isomorphic or adaptive items using AIG, students can become familiar with key elements and encounter more difficult items. In this regard, AIG is the simple but very efficient way of offering practice items to students.

To be more specific, the online homework system (also called Item Management System; IMS) which was developed for a formative assessment in 2015 was aimed to have following distinguished features from the previous version: First, this new online system loads AIG functionality. Second, teachers can search and select the assessment items from the pool (*i.e.*, online item bank) in order to make their customized assessment tool. Third, immediate results and feedback are given to students, and these results are presented in graphs or charts for easy understanding. Fourth, feedback is given to students based on the achievement standards. Finally, the AIG feature allows students to learn more about parental items using clone items. Detailed framework of the online system developed in 2015 is illustrated in Figure 1.

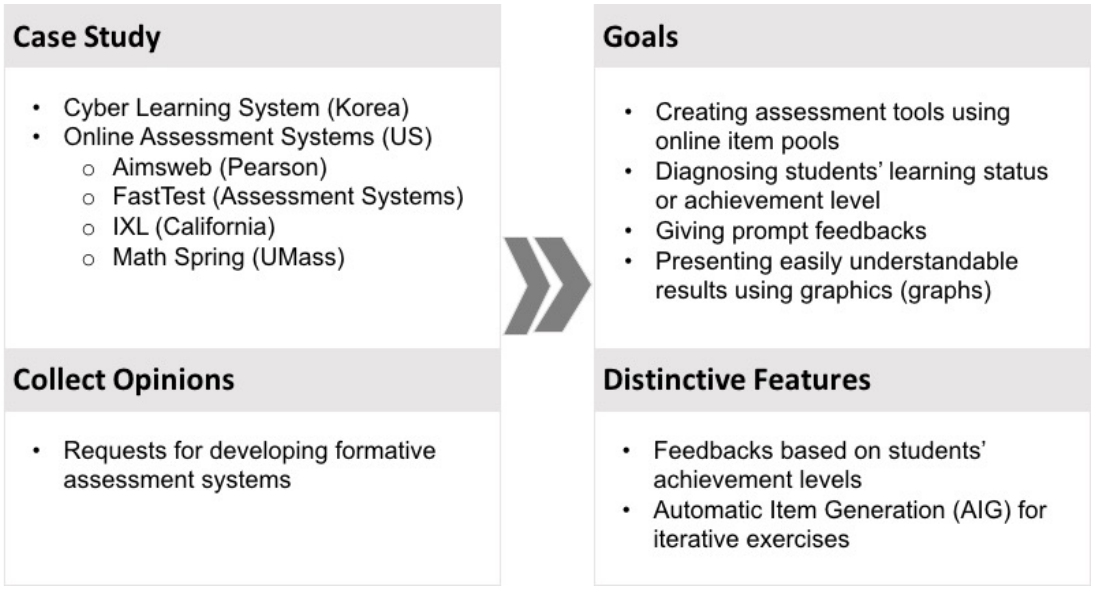

Figure 1. Features of Item Management System (IMS)

However, the claims for many of the potential benefits of AIG have not been adequately validated through the applications used in real education settings. A variety of educational applications using AIG functions should be developed and the results of validation studies for such systems should be accumulated sufficiently. These efforts will lead to the development of the appropriate and useful formative assessment system, and thus to the development of the education field.

1.1 Research Objectives

As we mentioned in the earlier section of this paper, it is in dire need of evaluating the utilities of AIG applications in practical educational settings. Based on these needs, the purposes of this research are addressed as follows:

(1) Introducing the benefits and challenges of AIG to properly assess AIG's utilities.

(2) Providing empirical evaluation results of AIG's utilities within a formative assessment system for Korean high school students and teachers.

(3) Providing theoretical and practical implications for AIG related research and practices.

This study has significances in that this is the first empirical study on evaluating AIG within a practical educational setting. In addition, AIG's theoretical and empirical assessment of this study will be useful to researchers and practitioners who wish to evaluate the benefits and problems of developing and/or applying AIG in educational services or products.

1.2 Theoretical Background

1.2.1 History of AIG

The history of AIG was characterized by diversity, that is, "*lack of cohesion among the*

contributors, and disconnectedness from current theory and research" (Gierl & Haladyna, 2013, p. 13). Yet, several milestones in the history of AIG can be identified as follows: 1) Prose-based AIG (Bormuch, 1970); 2) Facet theory (Guttman, 1953, 1959); 3) Item forms (Hively, 1974); 4) concept formation (Markle & Tiemann, 1970); and 5) the publication of Item Generation for Test Development (Irvine & Kyllonen, 2002).

In order to eliminate subjectivity in traditional item writing, Bormuth (1970) put forward his theory of item generation from prose by drawing on earlier research on cloze testing for reading comprehension and based on experiments of syntactic transformation. His theory was later refined and put into practice of item development (Finn, 1975; Roid & Haladyna, 1978). However, limited by the technology of the theory, items generated from prose seemed unable to measure higher-order cognitive ability. Other problems were also identified, and further improvement of the theory has not been reported.

Facet design (Guttman, 1953, 1959) is a proposal for domain-referenced testing, in which all possible tasks are identified in one domain and serve as the basis for testing. Mapping sentences are developed to define and put an order to the content in a domain and generate items, followed by empirical validation of its statistical structure. Every mapping sentence has fixed and variable parts and the latter is called a facet, which is how Facet Design got its name. Facet design allows for efficiency and objectivity in item writing, and also provides rich opportunity for diagnostic and formative assessment as well as computerized testing. However, some degree of subjectivity still exists in the development of mapping sentences.

The third milestone is called "Item Forms" (Hively, 1974), which generates items with a fixed syntactic structure and several variable elements. By putting constraints on the range of the replacement sets and adding conditions, a set of sentences can be defined and items can be automatically generated. This theory has the greatest applicability with items being quantitative in nature and items can be generated rapidly and abundantly and can be used for both formative and summative tests. However, manual work is still needed in developing item forms and usually a very large number of item forms are needed to define a domain.

As concept learning is one of the most important aspects in almost all fields of learning, interests in developing items measuring concepts surged in 1960s and 1970s (Markle & Tiemann, 1970; Anderson, 1972). Ways of generating items include paraphrasing concepts or generating examples and non-examples by manipulating the critical and variable attributes of a concept. The next milestone is the publication of Item Generation for Test Development (Irvine & Kyllonen, 2002). In this book, Irvine (2002) identified three measurement paradigms for AIG, which are R-models (dealing with achievement tests), L-models (dealing with timed testing), and D-models (dealing with predictive testing).

In the first step to generate an item, Bejar (2002) emphases a thorough analysis of the construct followed by the identification of task domains consistent with the construct. As a next step, "Item Models", which produce items with similar psychometric properties (called isomorphs) or variant difficulty levels, are developed. He also proposes two methods, the "Item Shells" and the mapping sentence, in response to practical needs. Item shells are from items with excellent psychometric properties in an item bank. By removing the content,

keeping the syntactic structure of these items, and adding different content, new items are created. Compared to the item model, the item shell is less constrained, gives more freedom in item generation, and also proves easy and effective to use especially for subject-matter experts who lack skills and experience in item writing (Gierl & Haladyna, 2012).

1.2.2 Applications of AIG

In the 21st century, the development of information and communication technology (ICT) has led to a rapid development of theories and practices of item generation to computer-based or computer-assisted item generation, the so-called Automate Item Generation (AIG). AIG has been used in a range of areas, such as K-12 subjects, psychological testing, and licensure/certification, etc. For examples in K-12, Choi, Kim, and Yoon (2014) developed a web-based AIG system and used the AIG system in various K-12 math applications: such as online workbook application (Choi, Kim, & Yoon, 2016a; Choi, Kang, Kim, Dardick, & Zhang, 2015), grade 6 paper version math workbook (Choi, Kim, & Yoon, 2016b) and grade 7 paper version math workbook (Choi, Kim, & Yoon, 2017) which are integrated with AIG-based system. Alves, Gierl, and Lai (2010) illustrated how to use AIG to create items for advanced placement (AP) biology. Yung and Choi (2018) also examined the potential benefits of AIG on various international assessments. In psychological domains, Bejar (1990) showed how to use computer algorithms to create item templates to measure mental rotation abilities while drawing on research results from cognitive science. Embreston (2002) presented an AIG-type generative system to generate abstract reasoning items based on cognitive theories. Alsubait, Parsia, and Sattler (2012) developed analogy multiple choice (MC) items from existing ontologies.

In the area of licensure and certification, Gierl, Lai, and Turner (2012) described how to use AIG to create multiple choice items for a medical licensure test in the content area of surgery. Karamanis, Ha, and Mitkov (2006) attempted to produce MC items from medical texts. In language related areas, Brown, Frishhoff, and Eskenazi (2005) made use of WordNet to automatically generate two types of English vocabulary assessment tasks: word bank and cloze question. Susuanti, Iida, and Tokunaga (2015) developed English vocabulary tests using TOEFL vocabulary question as a model. In grammar testing, Chen, Liou, and Chang (2006) introduced a method for semi-automatically generating grammar test items by applying Natural Language Processing (NLP) techniques. Perez-Beltrachini, Gardent, and Kruszewski (2012) also generated grammar exercises semi-automatically but for second language learners. Other researchers have focused on AIG for cloze tests. For example, Coniam (1997) described a process to automatically generate vocabulary cloze test items using word frequency data from an analyzed corpus. Liu, Wang, Gao, and Huang (2005) applied NLP techniques to algorithmically generate reading cloze items automatically. Goto, Kojiri, Watanabe, Iwata, and Yamada (2010) developed a system for automatically generating MC cloze questions from English texts.

1.2.3 Benefits of AIG beyond Mass Production

Although AIG is gaining in popularity, it is primarily known for its advantages in terms of production speed and unit price reduction of items (Choi & Li, 2016; Choi, 2017). However,

the benefits of AIG are not limited to mass production, and the benefits of this mass production must also be considered along with the cost and time required to deploy AIG (Choi, 2017). For example, the time and cost of training AIG item writers (or AIG template developers) and developing AIG templates should be fully considered and evaluated before adaptation of AIG in an assessment practice.

Choi (2018) also emphasizes the fact that many researchers have already argued that the benefits of AIG have been above and above mass production. In this regard, Choi analyzes the benefits and implications of AIG into two perspectives: theoretical and practical. For the theoretical perspectives beyond mass production, Choi illustrates the implications and benefits of AIG across the following points, with an emphasis on quality of assessment:

- Construct Preservation: AIG presents a new concept for traditional test security and offers a variety of options to address the threat of test security. Using AIG, test builders and administrators can effectively and differently handle test security concerns and improve instruction-assessment alignments via AIG.

- Construct Representation: AIG provides several options for further improving construct representation.

- Scientific Test Design: AIG plays a key role in recent scientific test design approaches. [*i.e.*, Evidence-centered Design (Mislevy, Almond, & Lukas, 2004), Cognitive Design System (Embretson, 1998) and Assessment Engineering (Luecht, 2014)].

- Item Calibration: AIG provides several strategies to improve item calibration and validation more efficiently.

- Equating and Linking: AIG can increase the efficiency of test equating and linking by allowing it to produce superior quality common items or tests.

- For AIG's *practical* implications and benefits of AIG, Choi (2017) also reviews and summarizes them across the following aspects:

- Item Management: AIG will provide efficiencies in storing, delivering, editing, and revising items. In particular, AIG provides an excellent option for creating and managing global assessment contents.

- Learner-centered Assessment: AIG can be used as a valuable tool to transform existing assessments into more learner-centered assessments such as diagnostic and/or formative assessment. AIG strengthens the links the skills to be measures with interpretations of test score; helps probe deeply into an examinee's weakness; is useful in diagnostic classification. And in formative assessment setting, AIG provides consistent, timely feedback, sufficiently large, and individualized feedback and diagnoses.

- Intelligent Assessment: We can utilize and fuse various ICT technologies (big data analysis, virtual or augmented reality, simulation based inferences, etc.) to develop existing assessments into more intelligent assessments beyond the traditional paper-pencil or computer-based assessment. A series of these theoretical and technical efforts can be

abbreviated as Assessment Engineering, and AIG plays a central role in building such intelligent system.

1.2.4 Benefits of AIG in Formative Assessment

In this section we will look more closely at what role AIG plays in formative assessment and how it can enhance the assessment. Formative assessment refers to assessment that is specifically intended to generate feedback on performance to improve and accelerate learning (Sadler, 1998). In contrast to summative assessment, which is usually administered at the end of a period of learning for the purpose of assessing the effectiveness of a program and documenting the progress an entire group of students make, formative assessment is typically administered during the process of learning and featured by constant feedback. Feedback is "*information about the gap between the actual level and the reference level of a system parameter which is used to alter the gap in some way*" (Ramaprasad, 1983, p. 4). As we will argue here, for an assessment to be formative and supportive for students' learning, three elements need to be incorporated in the process: 1) accurate diagnosis of students' present state of knowledge and skills measured; 2) immediate feedback provided to the learner and the teacher; 3) corresponding adjustment made in teaching and learning. Generally, after treatment, new assessments are needed to evaluate the effect of treatment or adjustment, and collect evidence for improvement in learning.

Formative assessments are usually carried out by instructors (*i.e.*, teachers). However, there are substantial challenges faced by teachers in implementing formative assessment (Choi, 2017). First, as response data or information about the degree of student learning is generally very diverse and complex. Therefore, teachers have to rely on their own observations and subjective judgments. Thus, a diagnosis that depends on the teacher's subject judgment may not maintain high level of accuracy and reliability. Consequently, feedback may not always be appropriate and/or effective enough to improve student's learning. Second, teachers may not be able to give timely feedback to every student especially when the class size is large. In most cases, the teachers are suffering from diverse and considerable tasks, and there is insufficient time for teachers to adequately diagnose and prescribe formative assessments for many students. Lastly, after the treatment and adjustment in learning are made, teachers need to design new tasks to evaluate the effects of teaching and learning. However, such tasks pose an even greater challenge to teachers, because such tasks "*must be sufficiently dissimilar from those previously attempted as learning exercises to test real achievement rather than memory and regurgitation ... they must also be similar enough to fall within the region that reasonably allows transfer or extended application of learning*" (Sadler, 1998, p. 81). These difficulties and challenges can be formidable for teachers to bring out the proved benefits of formative assessment for students' learning (Black & William, 1998; Choi, 2017). Here again, it is argued that AIG has the potential to help teachers address these challenges.

As mentioned above, with AIG-enabled assessment system, we can provide an efficient and reliable diagnosis of the strengths and weaknesses of individual students. This does not mean that teachers should waive the effort of devising their own formative tasks and rely solely on such tests. Instead, these assessments can complement the formative assessment tasks

designed by teachers and help provide a rapid and accurate picture of students' learning state. As we argued earlier, accurate and reliable information is important for subsequent treatment or remediation to be well-targeted and effective. It is clear that the diagnostic information can be provided to and used by both teachers and learners. Based on the information provided by the system, teachers can quickly assess students' learning needs and adjust their instructional plan to meet the needs of students. The system also let teachers have the option to give students more detailed feedback and suggestions for further actions or strategies students may take. Learners can use the information to reflect on their own learning process and make adjustments to their learning strategies.

After the treatment and remediation or intervention, tasks are needed to evaluate the effects of adjusted teaching and learning, and evidence needs to be collected to show if there is improvement in learning. Those tasks should be equivalent/consistent, *i.e.*, similar in difficulty and measure the same construct as those administered before the treatment and remediation; yet as different in forms as to test real improvement in learning instead of simple memory. This is not easy and can be very challenging especially when different learners need different sets of tasks and different amount of practice. AIG, using calibrated item templates, can generate a large number of isomorphic (same psychometric characteristics) items to meet individual students' needs (Choi, 2017).

More AIG-features can be integrated into the system to enable more effective feedback and support more tailored teaching and self-regulated learning. Sadler (1989) proposed the self-regulated learning that for real improvement in learning, students should be aware of the standards for learning and possible strategies for adjusting their learning and be given the opportunity to develop their own evaluative expertise. To achieve this purpose, detailed explanation or correct solutions can be given to students to enable them to evaluate their own work against the standards. Meanwhile, suggestions for effective learning strategies can be embedded in the automatic feedback.

However, the above-mentioned claims about the merits of AIG's formative assessment have not yet been sufficiently validated. Given the future-oriented nature of AIG and the increasing demand for formative assessment, a learner-centered assessment, empirical verification of the claims mentioned above is urgent. Researchers and practitioners who are interested in developing and/or applying AIG in educational services or products need more empirical evidence of evaluating the claims on AIG benefits on formative assessment in practical setting. Next part of this paper, we will provide an empirical evaluation results of AIG's utilities within a formative assessment system for Korean high school students and teachers.

2. Methods

The Item Management System (IMS) is an online formative assessment system used as a homework application, developed in the current study was designed for helping both teachers and students. For example, assessments given to students during the course work or at the end of each chapter could provide results and feedback which can identify the bottle-neck point in the learning process of students. Knowing the bottle-neck point can help teachers plan the best teaching strategies based on students' achievement levels or levels of understanding.

2.1 Development of Online Formative Assessment System

In sum, the online formative assessment system, which offers diagnosis and feedback based on the achievement levels, has four competences over the other standard formative assessments. First, three levels of feedback are available: item level, achievement level, and curriculum (*i.e.*, chapter) level. Second, students' learning strength and weakness are provided for each chapter over achievement levels. Third, the system includes mapping structures for achievement levels so that students can review previously unattained learning objectives. Finally, provide opportunities to deepen items using AIG technology as well as answers and explanations. CAFA AIG (Choi, Kim, & Yoon, 2014a) system was as AIG system for this application.

CAFA AIG (Choi, Kim, & Yoon, 2014a; http://AIG.CAFALab.com/) is web-based AIG system which it is built based on an assessment engineering framework. The system can provide nearly a limitless set of assessment items which are highly interactive and multi-media assessment items using several multi-media technologies, such as parameterized mathematical expression generation, parameterized chart/figure generation, or text-to-speech, etc (Choi, 2017). CAFA AIG also provides assessment services and functions for a variety of client-applications (*e.g.*, websites, mobile applications or eBbooks) as a cloud assessment platform. Each client application in the CAFA AIG system can use the AIG services to access the CAFA AIG server without having separate assessment items and services generated/developed by AIG technology (Choi, 2017). The IMS system and K-Math Workbook (a.k.a., CAFA SmartWorkbook; Choi, Kang, Kim, Dardick, & Zhang, 2015; Choi, Kim, & Yoon, 2016b; Choi, Kim, & Yoon, 2017) are examples of client applications using the CAFA AIG system as platform. Users who use these applications can use AIG services (such as using QR codes for immediate feedback) both online while they are connected to the server or offline by printing out workbook.

In the IMS system, the item explorer and the test builder are two basic functions that any online formative assessment system possesses. The online formative assessment system for this study has two updated functions compared to the previous version. Specifically, the online system in this study has AIG functionality so that students can have opportunities to go over items which they incorrectly answer. In addition, feedback could be more precisely described in that including diagnosis information as well as directions for future achievement in learning. These two additive functions are expected to expedite the self-learning strategies of students and behave as more beneficial feedback for students in learning.

IMS provides three levels of feedback: item level, achievement standard level, and unit level. The feedback made under the achievement standard levels diagnose students' learning standard into different levels (*i.e.*, high, middle, low, and under-achieved). The feedback from the unit level possesses information about the strength and weakness in learning. The utilization of a mapping structure in achievement standard allows to diagnose parts which need more exercise and review within the same school year, and even within the previous year of coursework.

2.2 Procedures

The IMS developed in this study can be delivered in two different formats as described in Figure 2. That is, teachers can choose the assessment format either being online (*e.g.*, computer) or offline (*i.e.*, paper-and-pencil) based on the given conditions at schools. Since the only difference between these two formats is a delivery option, the rest of assessment process remains the same. More specifically, every task, such as creating assessment tools, giving a test, rating, and providing feedback, under the online format is done using computers. On the other hand, a test under the offline format is given in a printed format to the students. After receiving the printed test, students mark their answers using computers, tablets, or smartphones. Answers are automatically rated, and results are given to students with feedback like the online format.

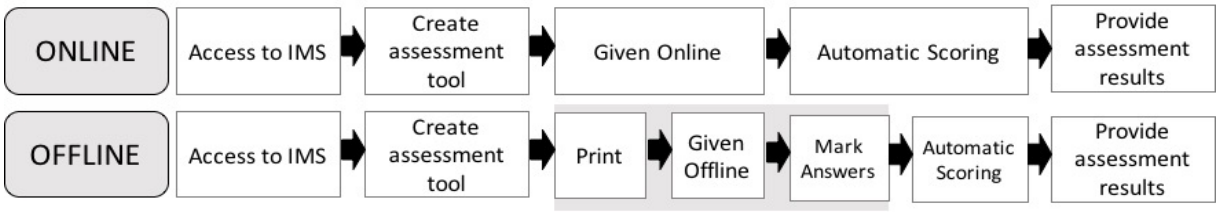

Figure 2. Flow charts for online and offline formats

When students click the button for explanations after submitting their answers, the screen shows the correct answers and explanations as well as providing practice problems using clone items of AIG. Figure 3 is the example of the scoring table. The first row contains the item number information with two distinguished colors on it: blue for the correctly answered item and orange for incorrectly answered item. The second row shows the correct answers for each item, and the third row encloses what students mark on the answer sheet. That is, when the numbers in the second row and third row for each column are corresponding, that item is answered correctly so that having blue color on the first row. The fourth row contains the icon ( ) for each cell which includes explanations for relevant items. Every cell in the fifth row has the icon ( ) for having exercise problems using clone items. More detail descriptions for the last two functions are given in Figures 4 and 5, respectively.

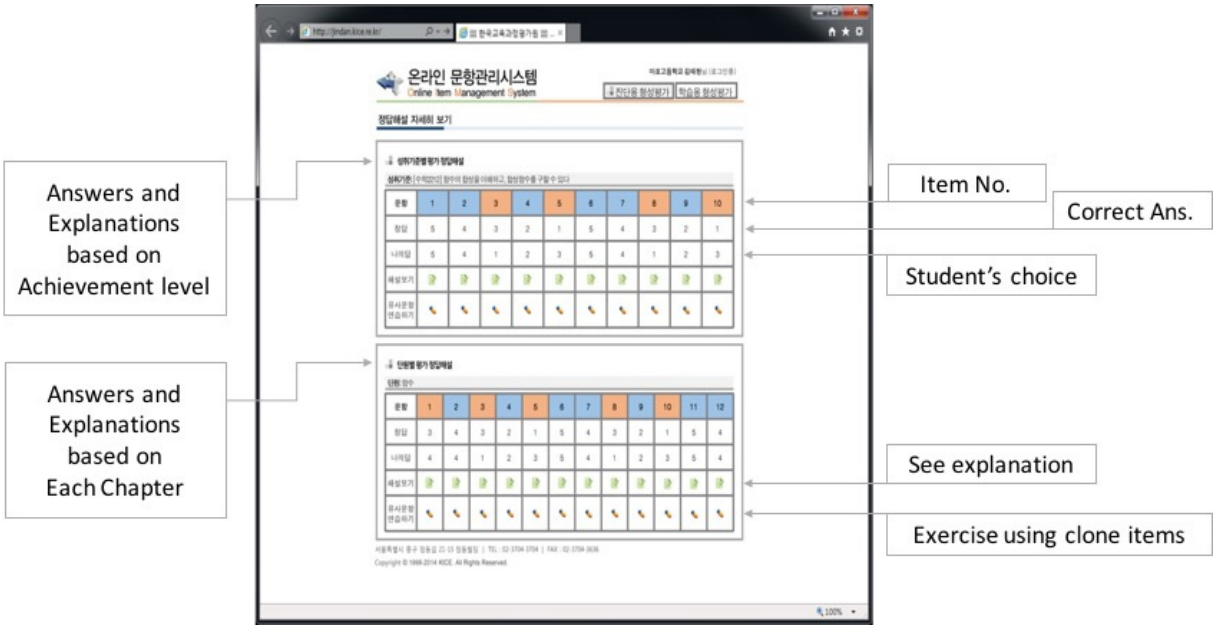

Figure 3. Actual screen shot shown to students

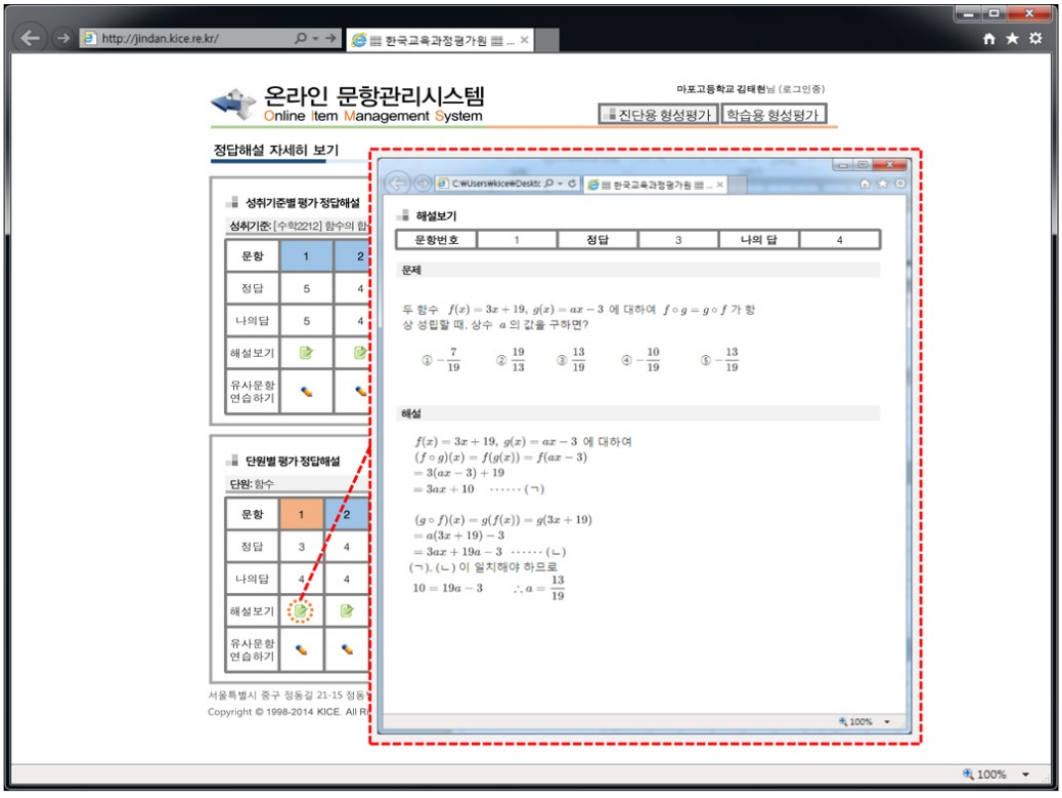

Figure 4. Actual screen shot when clicking "See Explanation ( )"

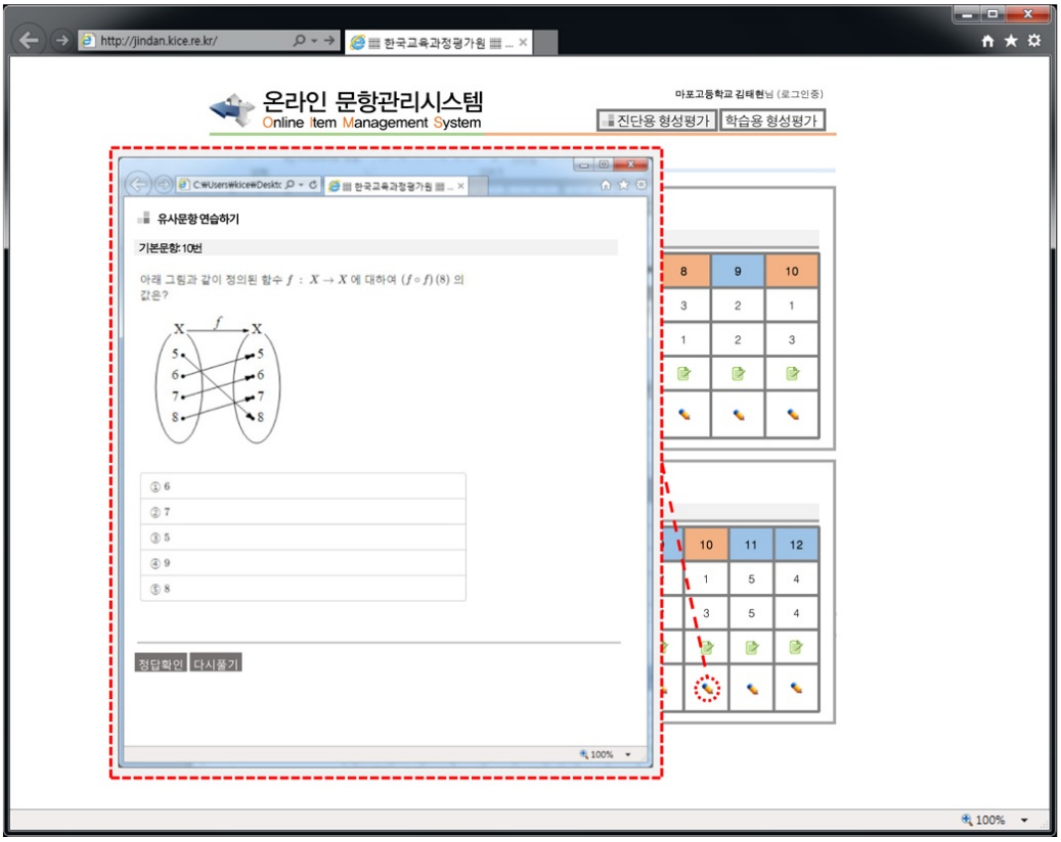

Figure 5. Actual screen shot when clicking "Exercise Clone Items ( )"

3. Results

The students were asked how helpful exercise problems were in learning mathematical functions. Fifty-two students (91.2%) were answered either "satisfied" or "very satisfied." The students were also asked which part of the IMS was most helpful in learning mathematical functions. Twenty-six students (45.6%) were selected "exercise problems using clone items", 16 students (28.1%) chose "explanations", and 13 students (22.8%) responded "diagnosis and feedback". The overall survey results are summarized in Figure 6 below.

Results of survey on Item Management System (IMS) (*N* of Respondents = 57)	
Convenience of IMS	Fifty students (87.7%) chose either 'agree' or 'completely agree' for their responses in asking "how convenient IMS is for the formative assessment".
Appropriateness of Feedbacks in IMS	Forty eight students (84.2%) chose either 'agree' or 'completely agree' for their responses in asking "whether feedbacks given are appropriate to understand the current achievement level in math".
Usefulness of Feedbacks in IMS	Fifty two student (91.2%) responded to either 'agree' or 'completely agree' in asking "whether assessment results (feedbacks) given are useful".
Satisfaction toward IMS	Overall satisfaction toward IMS was asked and 51 (89.5%) students answered either 'satisfied' or 'very satisfied'.
Most **helpful parts** in IMS (open-ended)	1. Iterative exercise using clone items of AIG (*N*=26, 45.6%) 2. Answers and explanations (*N*=16, 28.1%) 3. Diagnosis and help (*N*=13, 22.8%)

Figure 6. Survey results on example application of Item Management System (IMS)

Fifty-seven students were asked to take the survey related to using IMS (*i.e.*, IMS equivalently). The brief results of the survey are shown in Figure 7. Fifty (87.7%) students chose either 'Agree' or 'Strongly agree' for their responses in asking the convenience of using IMS in their actual formative assessment. The appropriateness of feedback was also asked and 48 (84.2%) students selected either 'Agree' or 'Strongly agree' in that survey question. When students were asked whether the assessment results and feedback given in IMS are useful, 49 (86.0%) students responded either 'Agree' or 'Strongly agree'. The overall satisfaction in experiencing IMS was also asked and 51 (89.5%) students either agreed or strongly agreed to that statement.

Other than Likert-scale survey questions, the open-ended question asking "what is the most helpful part in IMS?" was also given to the students. The following three things were ranked the top three of the all responses. First, twenty-six (45.6%) students thought that the iterative practice using clone items given by AIG is most helpful. Second, answers and explanations according to the achievement level were selected as the most helpful part in IMS for 16 (28.1%) students. Third, thirteen (22.8%) students answered that diagnosis and help session was most helpful in using IMS.

Results of survey on Item Management System (IMS) (*N* = 57)					
Survey Questions	**Strongly disagree**	**Disagree**	**Neutral**	**Agree**	**Strongly agree**
Based on your experience, was it convenient to use IMS for your formative assessment?	0 (0.0%)	1 (1.8%)	6 (10.5%)	22 (38.6%)	28 (49.1%)
Based on your experience, was the feedback given in IMS appropriate to understand your achievement level in math?	0 (0.0%)	1 (1.8%)	8 (14.0%)	25 (43.9%)	23 (40.4%)
Based on your experience, were the assessment results and feedback given in IMS useful?	0 (0.0%)	0 (0.0%)	8 (14.0%)	24 (42.1%)	25 (43.9%)
Based on your experience, were you satisfied with IMS overall?	0 (0.0%)	0 (0.0%)	5 (8.8%)	31 (54.4%)	20 (35.1%)

Figure 7. Survey results for IMS

Teachers also interviewed about overall satisfaction and usefulness of IMS, and asking for improvement plans if they were not satisfied with IMS. Overall, the results showed that teachers' responses toward IMS were promising and positive. Teachers said that information on student feedback and records in using IMS was particularly useful. This is because this information can be used in reference or recommendation letters that should be objective or unbiased. Teachers were also satisfied with prompt feedback features given to students as soon as they submitted answers to the assessment items under the online format.

On the other hand, the teachers identified some undesirable points and recommendations for better use of IMS. For example, students might suffer from difficulties in handling test items on the screen. This unfamiliarity can cause measurement issues, so it is necessary to use a more user-friendly service design or to improve other screen-related issues such as font size. In addition, there need some supporting strategies so that less able students can participate and intrigue more in using IMS. In other words, teachers concerned about the various possible approaches for students in different achievement levels. The automatic pop-up function for incorrectly answered items was also suggested for the simplicity in the online system itself.

In summary, the evaluation results showed that the various features of formative assessments based on AIG technology are highly promising in utilization for both students and teachers. In this study, we found that AIG has a variety of benefits when used in formative assessments, and that AIG-based functions (*e.g.*, online practice, adaptive diagnosis, and instant feedback, etc.) are well received by students and teachers.

4. Conclusion and Discussion

The individualized or tailored education, in which individual differences are taken into

account, requires some assessment activities that make it possible to comprehend the weakness and strength in students' learning. Based on the changed paradigm in the use of assessment (*i.e.*, a purpose of assessment), a dysfunctional aspect of assessment, such as trigging competitive atmosphere among students, has been disregarded, and more educational and positively functional aspects of assessment are being emphasized. In spite of this new trend in assessment, it is never easy for teachers to prepare sufficient and well-qualified formative assessments frequently and to do other school related tasks at the same time. In this respect, the assessment items developed at the national level rather than individual teacher level seems to be preferred.

Developing assessment items under the national level could embed broader aspects of items and control the quality of items more easily. This also could reduce the tasks and burden of a teacher so that they can more focus on teaching. For these reasons, establishing the infrastructure for online formative assessment system at the national level is necessary. An empirical evaluation of IMS which is based on AIG to the real school setting was conducted in order to verify the stableness and fineness of IMS focused on AIG feature.

The aforementioned IMS is expected to behave as the prototype in materializing formative assessments in real school settings as a homework system. Before making this system available for all schools, its stableness in major functions (especially AIG functionalities) and appropriateness in real school settings should be fully examined and evaluated. Therefore, this study was conducted as one example application of IMS for investigating its feasibility in practical usage focus on AIG utilities. In detail for research setting, two high schools in Seoul, South Korea, were selected for the IMS application. Fifty-seven students of the first grade in high school used IMS for two days. One chapter dealing with algebra was used for this application study. After using IMS, these students conducted a survey asking how satisfied IMS was.

The IMS used in this study is equipped with the item database (DB) like other standard item banks. Therefore, teachers can access the DB anytime to create their own assessment tools based on the needs. They can also use the assessment tools which are already composed and included in the DB. Teachers can assign a test to each student through the online, or distribute the printed version of the test to the students during the class. Regardless of the test format given to students, test results are generated in a short time, and both teachers and students can check the results. Test results could be checked via online, saved and even printed out.

Unlike other formative assessment systems, the IMS in this study is based on AIG technology as its specialty. Including AIG feature is mostly beneficial in that providing practice items by use of clone items in IMS. This feature was found to greatly contribute students' learning process. Students, as well as teachers, were quite satisfied with several AIG-based features of the system. In detail, other formative assessment systems are designed to primarily provide the information about students' achievement level at the most. However, knowing the achievement level is never enough to use assessments for learning beyond evaluation. The detailed and tailored feedback based on the achievement levels using AIG make the IMS be more useful for both students and teachers. Furthermore, having as many qualified items as

possible in the DB is a prerequisite condition for IMS to work well as a sound formative assessment system. For this purpose, AIG was adapted to the system and evaluated its effectiveness. Using the pre-validated and expertly developed AIG item template, the system can produce large amounts of quality items.

In practice, teachers frequently feel the burden of developing assessment items, grading and analyzing answer sheets, and providing acute feedback to students. In order to ensure a sound formative assessment in the current school environment, we must first reduce these overloaded tasks for the teacher. For this reason, this study proposed the use of online formative evaluation systems that included AIG functions and empirically evaluated their functions. The core part of IMS development is to have sufficient numbers of items in the item DB using AIG. The DB is not just the place where items are stored but where items should be searched and sorted based on teachers' needs, and controlled in their quality. The access to the DB also should not be complicated so that teachers could comprise items easily to make a test having appropriate characteristics, such as test difficulty. In addition, the information related to the achievement standard in the DB should be acute and reliable to make feedback useful.

Algorithms in generating clone items may vary depending on the subjects, though, we used CAFA AIG (Choi, Kim, & Yoon, 2014a) engine to generate high school level math items in this study. In Lai, Alves, and Gierl (2009) work, 64,260 clone items were generated from 34 parent items for various subjects (*e.g.*, Math, Literature, Science, and Social studies) of the 3rd, 6th, and 9th grades. Other AIG projects on math subject (Choi, Kim, & Yoon, 2016b; Choi, Kim, & Yoon, 2017) also showed more than two million unique math items can be generated by 350 item templates in 6th or 7th grade level. These studies also showed that AIG system could be used for other subject area beyond Math. For sure, it is hard to tell that AIG system automatically generates clone items in the equal level of quality as their parent items. However, not many differences were found between items generated by AIG system and those by human experts (Gierl & Lai, 2013). This implies that the integration of AIG system into IMS is efficient and beneficial in terms of time, effort, and money for item development.

This study aimed to develop and evaluate the IMS which is centered on students and customized to students' achievement level using AIG advancements. That is, specifying what their achievement level is not enough under the paradigm of "assessment for learning". Students should be provided with more detail feedback on what they are good at and what their weakness is. This feedback is better to include customized future directions for students based on their current achievement levels. It is not saying that the teacher-centered supplementary learning process is meaningless. Rather, it is better to say that the student-centered customized learning process may be more helpful and efficient in developing self-directed learning (SDL) strategies in students. The importance of developing SDL is that students can cope with their pace in learning, find motivations to learn, and develop a metacognition, which is known for a higher order thinking skill. For these customized learnings, the acute diagnosis about each student's learning status and their needs is necessary which are usually done by teachers.

This study demonstrated a great commitment to AIG's role in these new and forward-looking assessment applications using empirical evaluation in high school educational setting. Therefore, for both teacher and students, the utilization of IMS backed by AIG can be beneficial and promising in that the IMS provides acute, reliable, and unbiased assessment services and information. We hope that the usefulness of AIG will be further verified in various assessment applications, other subjects, and other grades in future studies.

References

Afzal, N., & Mitkov, R. (2014). Automatic generation of multiple choice questions using dependency-based semantic relations. *Soft Computing, 18*(7), 1269-1281. https://doi.org/10.1007/s00500-013-1141-4

Agarwal, M., Shah, R., & Mannem, P. (2011). Automatic question generation using discourse cues. *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1-9). Association for Computational Linguistics.

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

Alsubait, T., Parsia, B., & Sattler, U. (2012). Automatic generation of analogy questions for student assessment: An ontology-based approach. *Research in Learning Technology, 20*(5), 95-101. https://doi.org/10.3402/rlt.v20i0.19198

Alves, C. B., Gierl, M. J., & Lai, H. (2010). *Using automated item generation to promote principled test design and development*. In Annual Meeting of the American Educational Research Association, Denver, CO, USA.

Arendasy, M. E., & Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and Individual Differences, 22*(1), 112-117. https://doi.org/10.1016/j.lindif.2011.11.005

Bejar, I. I. (1990). A generative analysis of a three-dimensional spatial task. *Applied Psychological Measurement, 14*(3), 237-245. https://doi.org/10.1177/014662169001400302

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. In N. Frederiksson, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 323-359). Mahwah, NJ: Lawrence Erlbaum Associates.

Bhatia, V. (2004). *Worlds of written discourse: A genre-based view*. London: Continuum.

Bormuth, J. R. (1970). *On a theory of achievement test items*. Chicago, IL: University of Chicago Press.

Brown, J. C., Frishkoff, G. A., & Eskenazi, M. (2005). *Automatic question generation for vocabulary assessment*. Paper presented at the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP), Vancouver, Canada. https://doi.org/10.3115/1220575.1220678

Cai, Z., Rus, V., Kim, H. J. J., Susarla, S. C., Karnam, P., & Graesser, A. C. (2006). *Nlgml: A*

markup language for question generation (Vol. 2006, No. 1, pp. 2747-2752). In World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education.

Chen, C. Y., Liou, H. C., & Chang, J. S. (2006). Fast: an automatic generation system for grammar tests. *Proceedings of the COLING/ACL on Interactive presentation sessions* (pp. 1-4). Association for Computational Linguistics. https://doi.org/10.3115/1225403.1225404

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.

Choi, J. (2018). Roles and Impacts of Automatic Item Generation on Assessment Research, Practice, and Policy. In S. Swayze & V. Ford (Eds.), *Innovative Applications of Knowledge Discovery and Information Resources Management*. PA: IGI Global. https://doi.org/10.4018/978-1-5225-5829-3

Choi, J., & Li, L. (2016). *Automatic item generation: Beyond cost efficiency*. Presented at Annual Meeting of Korean-American Educational Researchers Association, Washington, DC.

Choi, J., Kang, M., Kim, N., Dardick, W., & Zhang, X. (2015). A smart way of coping with common core challenges—Introduction to CAFA SmartWorkbook. *Journal of Educational Issues, 1*(2), 70-89. https://doi.org/10.5296/jei.vli2.8381

Choi, J., Kim, S., & Yoon, K. (2014). *CAFA Automatic Item Generation (v. 1.0 Beta; http://aig.cafalab.com/) System: Computer Adaptive Formative Assessment Client Application for Automatic Item Generation* [Computer System]. CAFA Lab, Inc.

Choi, J., Kim, S., & Yoon, K. (2016a). *K-Math Workbook (v. 1.0) System: Computer Adaptive Formative Assessment Client Application for Common Core Math Workbook* [Computer System]. CAFA Lab, Inc.

Choi, J., Kim, S., & Yoon, K. (2016b). *K-Math Workbook Grade 6*. Clarksville, MD: CAFA Lab, Inc.

Choi, J., Kim, S., & Yoon, K. (2017). *K-Math Workbook Grade 7*. Clarksville, MD: CAFA Lab, Inc.

Coniam, D. (1997). A preliminary inquiry into using corpus word frequency data in the automatic generation of English language cloze tests. *Calico Journal, 14*, 15-34. https://doi.org/10.1558/cj.v14i2-4.15-33

Cubric, M., & Tosic, M. (2011). Towards automatic generation of e-assessment using semantic web technologies. *International Journal of E-Assessment*.

Curto, S., Mendes, A. C., & Coheur, L. (2011). Exploring linguistically-rich patterns for question generation. *Proceedings of the UCNLG+Eval: Language Generation and Evaluation Workshop* (pp. 33-38).

Davis, F. (1968). Research in comprehension in reading. *Reading Research Quarterly, 7*, 628-678. https://doi.org/10.2307/747108

Downing, S. M. (2006). Twelve steps for effective test development. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-25). Mahwah: Laurence Erlbaum.

Drasgow, F., Luecht, R. M., & Bennett, R. E. (2006). Technology and testing. In S. H. Irvine, & P. C. Kyllonen (Eds.), *Educational measurement* (4th ed., pp. 471-516). Washington, DC: American Council on Education.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3*, 300-396. https://doi.org/10.1037/1082-989X.3.3.380

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Erlbaum.

Gierl, M. J., & Haladyna, T. M. (2013). Automatic Item generation: An introduction. In M. J. Gierl, & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 3-12) Routledge.

Gierl, M. J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education, 46*, 757-765. https://doi.org/10.1111/j.1365-2923.2012.04289.x

Golestanirad, S. (2015). *Utilization of common human queries in ranking automatically generated questions* (Doctoral dissertation, Department of Mathematics and Computer Science, University of Lethbridge, Lethbridge, Alta).

Goto, T., Kojiri, T., Watanabe, T., Iwata, T., & Yamada, T. (2010). Automatic generation system of multiple-choice cloze questions and its evaluation. *Knowledge Management and E-Learning, 2*, 210-224.

Grabe, W., & Jiang, X. (2014). Reading assessment. In A. Kunnan (Ed.), *Companion to language assessment* (Vol. 1, pp. 185-201). Malden, MA: Blackwell.

Haladyna, T. M., & Gierl, M. J. (2013). Obstacles for automatic item generation. In M. J. Gierl, & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 231-239). Routledge.

Heilman, M., & Smith, N. A. (2010). Extracting simplified statements for factual question generation. *Proceedings of QG2010: The Third Workshop on Question Generation* (pp. 11-20).

Heilman, M., & Smith, N. A. (2010, June). *Good question! Statistical ranking for question generation* (pp. 609-617). In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing, 2*(2), 127-160. https://doi.org/10.1007/BF00401799

Irvine, S. H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Kalady, S., Elikkottil, A., & Das, R. (2010). Natural language question generation using syntax and keywords. *Proceedings of QG2010: The Third Workshop on Question Generation* (pp. 1-10).

Karamanis, N., Ha, L. A., & Mitkov, R. (2006). *Generating multiple-choice test items from medical text: A pilot study*. Paper presented at the Fourth International Conference Natural Language Generation, Sydney, Australia. https://doi.org/10.3115/1706269.1706291

Kintsch, W. (2004). The construction-integration modelof test comprehension and its implications. In R. B. Ruddell, & N. J. Unrau (Eds.), *Theoretical Models and Processes of Reading* (5th ed., pp. 1270-1328). Newark, DE: International Reading Association.

Lai, H., Alves, C., & Gierl, M. (2009). Using automatic item generation to address item demands for CAT. *Proceedings of the 2009 GMAC Conference on Computerized Adaptive Testing*.

Liu, C. L., Wang, C. H., Gao, Z. M., & Huang, S. M. (2005). Applications of lexical information for algorithmically composing multiple-choice cloze items. *Proceedings of the second workshop on Building Educational Applications Using NLP* (pp. 1-8). Association for Computational Linguistics. https://doi.org/10.3115/1609829.1609830

Luecht, R. (2014). Assessment Engineering Task Model Maps, Task Models and Templates as a New Way to Develop and Implement Test Specifications. *Journal of Applied Testing Technology, 1*(1), 1-38. Retrieved from http://www.jattjournal.com/index.php/atp/article/view/45254

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

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). *A Brief Introduction to Evidence-Centered Design* (CSE Report 632). CA: Center for Research on Evaluation, Standards, and Student Testing.

Mitkov, R., Ha, L. A., & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering, 12*(2), 177-194. https://doi.org/10.1017/S1351324906004177

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys, 41*(2), 1-69. https://doi.org/10.1145/1459352.1459355

OECD, & CERI. (2008). *Assessment for learning: Formative assessment*. Paper from the OECD/CERI International Conference "Learning in the 21st Century: Research, Innovation and Policy". Retrieved from http://www.oecd.org/site/educeri21st/40600533.pdf

Papasalouros, A., Kanaris, K., & Kotis, K. (2008). Automatic generation of multiple choice

questions from domain ontologies. In M. Baptista Nunes, & M. McPherson (Eds.), *E-Learning* (pp. 427-434). IADIS.

Perez-Beltrachini, L., Gardent, C., & Kruszewski, G. (2012). *Generating Grammar Exercises* (pp.147-157). The 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACLHLT Workshop 2012, Montreal, Canada.

Prasad, R., & Joshi, A. (2008). A discourse-based approach to generating why-questions from texts. *Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge* (pp. 1-3).

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile framework*. Paper presented at the 4th North American Conference on Adolescent/Adult Literacy, February 1996, Washington, D.C.

Susanti, Y., Iida, R., & Tokunaga, T. (2015). Automatic Generation of English Vocabulary Tests. *Proceedings of the 7th International Conference on Computer Supported Education (CSEDU 2015)* (pp. 77-78). https://doi.org/10.5220/0005437200770087

Wolfe, J. H. (1976). Automatic question generation from text-an aid to independent study. *ACM SIGCSE Bulletin, 8*(1), 104-112. https://doi.org/10.1145/952989.803459

Yang, Y.-C., Yang, J.-F., Chang, J.-M., & Chang, J. S. (2006). *Development of a Computer Assisted Reading Comprehension Test*. In International Conference on English Instruction and Assessment.

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

Young, D. & Choi, J. (2018). Information and Computer Technologies for Improving International Assessment. In S. Swayze, & V. Ford (Eds.), *Innovative Applications of Knowledge Discovery and Information Resources Management*. PA: IGI Global. https://doi.org/10.4018/978-1-5225-5829-3

Copyright Disclaimer