Educational Technology International
2022, Vol. 23, No 2, 157-181

https://doi.org/10.23095/ETI.2022.23.2.157

Application of AIG Implemented within CLASS Software for Generating Cognitive Test Item Models

Seungyeon SA
Yonsei University
Korea

Hyun Suk RYOO
University of Virginia
USA

Ji Hoon RYOO*
Yonsei University
Korea

Scale scores for cognitive domains have been used as an important indicator for both academic achievement and clinical diagnosis. For example, in education, Cognitive Abilities Test (CogAT) has been used to measure student's capability in academic learning. In a clinical setting, Cognitive Impairment Screening Test utilizes items measuring cognitive ability as a dementia screening test. We demonstrated a procedure of generating cognitive ability test items similar as in CogAT but the theory associated with the generation is totally different. When creating cognitive test items, we applied automatic item generation (AIG) that reduces errors in predictions of cognitive ability but attains higher reliability. We selected two cognitive ability test items, categorized as a time estimation item for measuring quantitative reasoning and a paper-folding item for measuring visualization. As CogAT has widely used as a cognitive measurement test, developing an AIG-based cognitive test items will greatly contribute to education field. Since CLASS is the only LMS including AIG technology, we used it for the AIG software to construct item models. The purpose of this study is to demonstrate the item generation process using AIG implemented within CLASS, along with proving quantitative and qualitative strengths of AIG. In result, we confirmed that more than 10,000 items could be made by a single item model in the quantitative aspect and the validity of items could be assured by the procedure based on ECD and AE in the qualitative aspect. This reliable item generation process based on item models would be the key of developing accurate cognitive measurement tests.

Keywords : Automatic item generation, CLASS, Cognitive test, Item model, Validity, CogAT

* Corresponding author: Department of Education, Yonsei University, ryoox001@yonsei.ac.kr

Introduction

Scale scores for cognitive domains have been used as an important indicator for both academic achievement and clinical diagnosis. In education field of United States, Cognitive Abilities Test (CogAT; Riverside Insights, 2022) has been used to measure student's capability in academic learning that helps educators understand their students' readiness. According to Warnimont (2010)'s findings, CogAT can be used to predict academic achievement while supporting the importance of data-driven decision-making. In clinical settings, cognitive ability is often repeatedly measured to confirm the effectiveness of treatment or appropriate time for a treatment (Stanek et al., 2011; Jutten et al., 2021). Cognitive Impairment Screening Test (CIST), the dementia screening test developed by Ministry of Health and Welfare of Korea, utilizes items measuring cognitive ability, categorized as executive functions such as identifying the sequences of shapes or perceiving a translation and a language function such as naming objects given.

In addition to the role of cognitive test in both academic and clinical settings, it is also important to make items of cognitive ability tests reliable as well as to make sure the underlying theory of cognitive test. In cognition, the Cattell-Horn-Carroll theory (CHC theory; Bryan & Mayer, 2020; Flanagan et al., 2013; McGrew, 2009; Schneider & McGrew, 2012; Schneider & McGrew; 2018) is a well-known theory synthesizing a variety of discussion on cognition. For example, CogAT as well as items that we created in this study are based on the CHC theory, which is also telling that every item in the current study has been verified the validity. However, unlike the items in CogAT, the item instances in our study have a fundamental difference in the procedure of generating items. We demonstrated a procedure of generating items that is called as automatic item generation (AIG; Drasgow et al., 2006; Embretson & Yang, 2006; Gierl & Haladyna, 2012; Gierl et al., 2021; Irvine & Kyllonen, 2002).

This main difference tells that AIG can reduce errors in predictions of cognitive ability and attain higher reliability. AIG is an innovative technology that automatically generates items from an item model that is coded (or digitized). By using AIG, it would be possible to prevent bias from practice effects or cheating as test forms consist of isomorphic, different-looking items. Goldberg et al. (2015) described practice effects as “an obvious disturbance factor in clinical trials.” It is noted that practice effects derived from repetitive performance in cognitive assessment result in a gradual improvement of subject’s performing (Wesnes & Pincock, 2002). It can be easily understood that AIG can remove these practice effects by a single item model in that it generates numerous isomorphic, different-looking items. Also, as each subject can be asked to respond to different item instances from a same item model, AIG can prevent cheating in tests. Reliability provides the information about a measurement result of how much it is stable and consistent in repetitive trials (Carmines & Zeller, 1979; Taherdoost, 2016). The items made by AIG have the higher reliability because the item model generation process is fundamentally related to examinees’ process of thinking which is represented with cognitive model. To explain it in a practical way, when content experts, who are professionals in the field which items belong to, constructing an item model, they control conditions of the model by using parameters, a coded variable that affects the characteristics of knowledge, skills, and ability to be measured in an item (Mislevy et al., 2003). This process, which will be explained more concretely through the application in the current item model instances, guarantees the reliability of items.

We have chosen Collaborative Learning Analytics Software Service (CLASS; CLASS, 2022) for the AIG software that is the only learning management system (LMS) including AIG technology in Korea. CLASS has many other features such as administering online tests, automatic scoring by using item response theory, but we focused on the feature of generating cognitive items in this study. CLASS is developed based on the two frameworks: evidence-centered design (ECD; Mislevy et al., 2003; Mislevy et al., 2004) and assessment engineering (AE; Gierl & Haladyna, 2012). Especially, three models among six models of conceptual assessment framework layer of ECD (Behrens et al., 2010; Mislevy et al., 2004) and three models of AE are the foundation of the item generation process in CLASS, which will be briefly discussed in the Theoretical Backgrounds section.

This study aims to demonstrate the item generation process using AIG implemented within CLASS, which can reduce problems of cognitive tests (e.g., bias from practice effects and cheating) and improve the reliability of items. The process of developing item models, (a) a time estimation item for measuring quantitative reasoning and (b) a paper-folding item for measuring visualization, were described in detail along with proving quantitative and qualitative strengths of AIG.

Theoretical Backgrounds

Automatic Item Generation (AIG)

Automatic item generation (AIG) is the process of using an item model to generate statistically calibrated items (Gierl & Haladyna, 2012), where statistical calibration is defined as the process quantifying the scale of a measuring instrument (Osborne, 1991). In other words, regarding to AIG, 'statistically calibrated items' means conceptually calibrated items with the process of generating item. The direction of the item generation process of AIG is not confirming the effectiveness and difficulty of items after producing them, but rather solidifying before doing so. Determining the effectiveness and difficulty of items in advance becomes possible by constructing an item model, defining variables related to the knowledge, skills, and abilities to be measured (Embretson, 1998).

Traditional tests designed in the form of writing items on paper and solving them with a pencil was referred as a paper-pencil testing design (Gierl & Haladyna, 2012). Through the development of computer technology, it was possible to design tests by making items manually with computers and printing them on paper to conduct test. Nevertheless, prior to the emergence of AIG, mechanism of writing items utilizing computer still did not differ significantly from the paper-pencil testing design. By the time of booming computer technology in assessment around 2000, completely differentiated new test design approaches, ECD and AE providing a theoretical foundation of AIG, have emerged. The emergence of these crucial foundations has geared up true digitization on item writing process. Traditional item generation methods such as using paper and pencils may easily create items, but its efficiency would be greatly reduced when a vast number of items are required to evaluate personal trait through repeated measurements. On the other hand, AIG technology gives the way to create massive amounts of items in a short period of time when an item model is generated. However, the advantage of AIG is not limited to its quantity of item generation. Using AIG, it becomes possible to create isomorphic items that are impossible with traditional item generation methods (Gierl & Haladyna, 2012). This became possible by using digitized item models. Furthermore, the validity of items is also guaranteed through the item model because the item model is created based on the process of cognition (Embretson, 1998). Gierl and Haladyna (2012) mention that item validation seeks evidence that each task/item does what it is supposed to do, thus any test is a representative sample from CHC-theory domain, each item should be proven to measure desired content with a predictable cognitive demand.

Mislevy et al. (2003) present the process of developing a digitized item model based on evidence, a variable that affects the characteristics of knowledge, skills, and ability to be measured in an item. The approach is called as evidence-centered (assessment) design (ECD). The items generated through the item model have the same degree of cognitive complexity (Embretson & Yang, 2006), which makes it possible to create numerous isomorphic items. Here, the isomorphism implies not only appearances of items as a question given but also degree of the item characteristics such as item difficulty. Behrens et al. (2010) presents five layers in ECD: domain analysis, domain modeling, conceptual assessment framework, assessment implementation, and assessment delivery. The description of each layer was briefly described below.

After analyzing an object (or a construct) to be measured in domain analysis, item writer represents variables affecting the object as evidence based on this analysis in domain modeling. Conceptual assessment framework is the process of constructing assessment models. In assessment implementation, students' test results are evaluated with models built in conceptual assessment framework. And then, in assessment delivery, students check the assessment of the test results and receive feedback.

The layer directly corresponding to AIG is conceptual assessment framework, consisting of six models: student model, evidence model, task model, assembly model, presentation model, and delivery system model. Especially, student model, evidence model, and task model among these models are directly connected with the item generation process using AIG. Thus, the realization of AIG within CLASS is based on these three models of conceptual assessment framework in ECD. The six models and the goals of each model in question format defined by Mislevy et al. (2004) are presented in <Table 1>.

Table 1
Six models of conceptual assessment framework and the goals of each model (Mislevy et al., 2004)

Model	Goal
Student model	▫ What are we measuring?
Evidence model	▫ How do we measure it?
Task model	▫ Where do we measure it?
Assembly model	▫ How much do we need to measure?
Presentation model	▫ How does it look?
Delivery system model	▫ Putting it all together.

Gierl and Haladyna (2012) defined AE as the use of engineering-based principles and technical processes for test design and development. AE consists of three models: construct map, task model, and template. Construct map is a cognitive-based model for test performance, which means the knowledge, skills, and abilities required to solve the task. The second model is a task model, specifically describing the knowledge and skills specified in the construct map. The last model is template that outlines the structure, content, limitations, and so on, required to produce items for measuring specific content defined in task model. Items generated through one template are guaranteed the same difficulty. According to Gierl and Haladyna (2012), by systematically manipulating the parameters during constructing a template, large numbers of items which are intended to measure content with comparable psychometric characteristics (e.g., similar difficulty levels), then the generated items are called isomorphs (Irvine & Kyllonen, 2002).

It can be simply understandable that three models of conceptual assessment framework in ECD and three models of AE explains same concepts in different words. For instance, the connotative meanings of student model in ECD and construct map in AE are almost same. Also, evidence model in ECD and task model in AE and task model in ECD and template in AE imply same meanings. Likewise, this study uses terms in ECD, student model, evidence model, and task model which have corresponding meanings with construct map, task model, and template in AE, to introduce AIG. In this study, we give shape to these frameworks by presenting the step-by-step process of AIG.

Cognitive Abilities Test (CogAT)

CogAT is a well-known education measure of cognitive ability, commonly used in Seattle, Dallas, Baltimore, Atlanta, North Carolina, South Carolina, Washington, D.C., Chicago, Minneapolis, Houston, San Antonio, and so on for admission to gifted programs or distinguishing below grade level students in schools. CogAT is more useful than Woodcock Johnson's test because they are capable of collective testing and have the potential to expand worldwide as a non-verbal test tool without language constraints.

CogAT consists of verbal, quantitative, and nonverbal/spatial sections to evaluate students' sequential, inductive, and quantitative reasoning abilities over K-12 grades (Thompson, 2011). Verbal battery measures the ability to infer through language data and flexibility, fluency, and adaptability when solving language problems. Quantitative battery measures quantitative reasoning ability, flexibility and fluency in quantitative symbols and concepts, and systematization, structuring, and meaning-giving ability for an ordered set of numbers and mathematical symbols. Finally, nonverbal battery measures inference using geometric shapes and figures. For example, students should present and implement strategies so that they can solve certain problems and perform successfully. According to Riverside Insights (2022), these reasoning skills can be used as good variables to predict good performance in school. According to Warnimont (2010)'s findings, CogAT can also be used to predict academic achievement while supporting the importance of data-driven decision-making.

As CogAT is certified and has widely used as a cognitive measurement test, developing an AIG-based cognitive measure has greatly contributed to education field. We have chosen two cognitive ability measurement items belonging to CogAT test, categorized as quantitative reasoning in fluid reasoning and visualization in visual processing of the CHC theory. Although there are many other items in CogAT, the reason why we've chosen these items is these are two of the most complicated items to build item models.

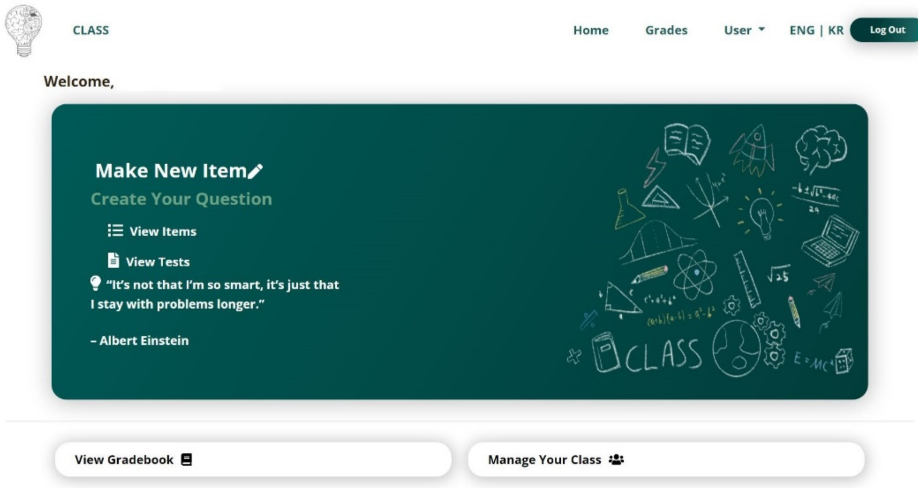
Development Methods

Collaborative Learning Analytics Software Service (CLASS)

CLASS is a platform for Learning Analysis as an LMS with four key features: (1) production of items and tests using AIG, (2) measurement using item response theory, (3) assessment system with adaptative test for learner's self-learning, and (4) cooperative assessment system for teachers. In this study, we only concentrate on the first characteristics, production of items and tests using AIG constructing item instances of cognitive measurement.

CLASS reflects all three main points that Gierl and Haladyna (2012) pointed out about AIG. The first point is that there is a need for someone to create expressive and generative item models. In that respect, CLASS has been already shown by being used in not only cognitive measurement but also college statistics and mathematics courses with items generated by AIG. The easy-to-use interface in CLASS would promote the usage in other fields to create expressive and generative item models. The second point is that it should be possible to implement the process of generating items with AIG. CLASS has built an item model generation page with an intuitive interface, and furthermore, it allows to check various example items to be generated while writing the item model using different seed numbers and preview functions. The third point is that the generated item model should be able to be stored. CLASS allows item models to be stored in personal item list, and furthermore, they can be registered on the common item list after going through the expert approval procedure. Moreover, as an LMS, CLASS allows teachers to immediately organize a test with items made by AIG and administer the test to students. It is very easy to access through the website, and it supports both Korean and English, allowing users to choose a language familiar to them.

The page generating an item model in CLASS can be accessed through two ways. As shown in Figure 1, 'View Items' allows to check the common item list to use or modify the items in it and by clicking 'Make New Item', a completely new item can be developed directly, stored in the personal item list of 'View Items'. These items written belonging to those lists can be easily combined into a test. More information about LMS and CLASS can be found in Park et al. (2022)'s paper, which is out of scope in the current paper.



"It's not that I'm so smart, it's just that I stay with problems longer."

- Albert Einstein

Figure 1. Main page of CLASS.

Item Generation Procedure Using Automatic Item Generation

Appling ECD and AE framework, CLASS has its own item generation procedure. Student model in ECD aims to define variables related to the knowledge, skills, and abilities to be measured (Embretson, 1998). Construct map in AE, a cognitive-based model for test performance (Gierl & Haladyna, 2012), has the corresponding meaning with the student model. In the example of the time estimation (belonging to quantitative reasoning) of the current study, student model (or construct map in AE) implies starting hour, starting minute, and time increase sequence parameters, which are the variables that affect the time estimation.

Evidence model (or task model in AE) provides a detailed explanation of how to update the variables defined in the student model. To update the time estimation variable, examinees should know a concept of the hour, the minute, and the calculation ability of time increasing. In this step, an item writer should consider examinees' process of thinking to solve the item by visualizing it using cognitive model. By constructing cognitive model, the item writer can decide which factors in the item should be parameters and which can be the worthwhile distractors. The example of the cognitive model of the time estimation item instance is shown in Figure 2.

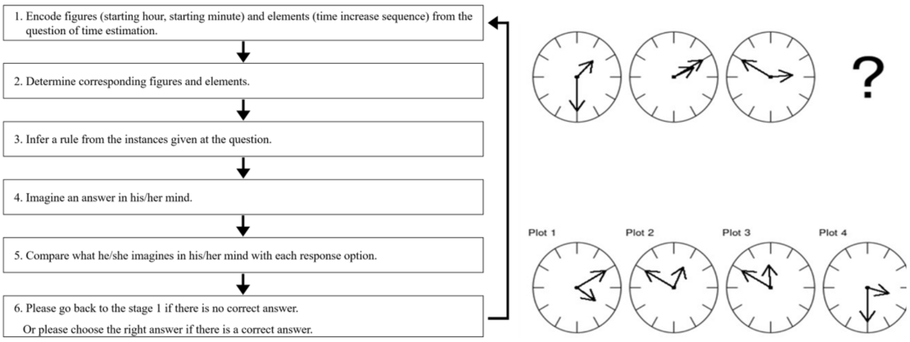
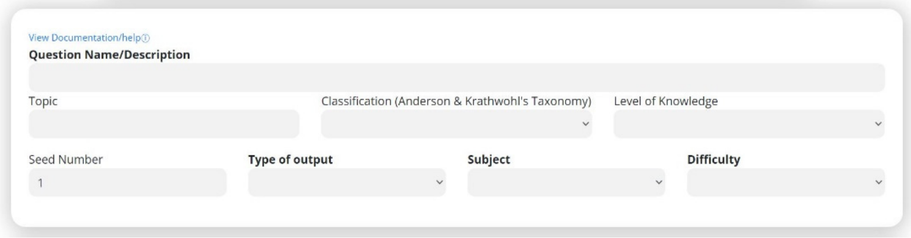

1. Encode figures (starting hour, starting minute) and elements (time increase sequence) from the question of time estimation.
2. Determine corresponding figures and elements.
3. Infer a rule from the instances given at the question.
4. Imagine an answer in his/her mind.
5. Compare what he/she imagines in his/her mind with each response option.
6. Please go back to the stage 1 if there is no correct answer.
Or please choose the right answer if there is a correct answer.
?
Plot 1
Plot 2
Plot 3
Plot 4

Figure 2. The cognitive model of the time estimation item instance.

Lastly, task model (or template in AE) aims to establish a real structure of the item model. Building the structure of the model through the task model will create numbers of items of the same difficulty. We use CLASS to show this third model by presenting the procedure about building the item model in detail in the Development Result section.

In the item model generation page, there is an item information section shown in Figure 3. In 'Question Name/Description', the name of the item and the description of the item can be entered. As in the case of this study, 'Time Estimation' and 'Paper-folding' is entered. In 'Topic', a topic related to the item should be entered, for example, 'Quantitative Reasoning' and 'Visualization' in this study. Through 'Classification', item writer assigns the item into Anderson and Krathwohl (2001)'s taxonomy, cognitive process dimension and knowledge dimension, that expanded Bloom's educational goal classification system (Bloom, 1956). Cognitive process dimensions are listed in the 'Classification' tab and the knowledge dimension is listed in the 'Level of Knowledge' tab with the dropdown menu of F-Factual Knowledge, Co-Conceptual Knowledge, P-Procedural Knowledge, and M-Meta Knowledge (Anderson & Krathwohl, 2001). CLASS provides random selection on parameters created by using a random seed, which presents isomorphic items generated in one item model differently for each student and each repeated trial. An item writer can check all items created by an item model be simply changing a number in 'Seed Number'. As for the 'Type of Output', one of equation, plot, cognition, and R-Markdown (RMD) can be selected, so we select plot for the item model instances in this study. 'Subject' option provides math, statistics, computer science, physics, and cognition so that the field of an item model can be specified. Finally, 'Difficulty' is the only space among all the components that cannot be entered by item generators. When an item model is first produced, it is displayed as 'Too New to Rate', and then it will be estimated based on the students' item responses.



View Documentation/helpⓘ

Question Name/Description

Topic

Classification (Anderson & Krathwohl's Taxonomy)

Level of Knowledge

Seed Number: 1

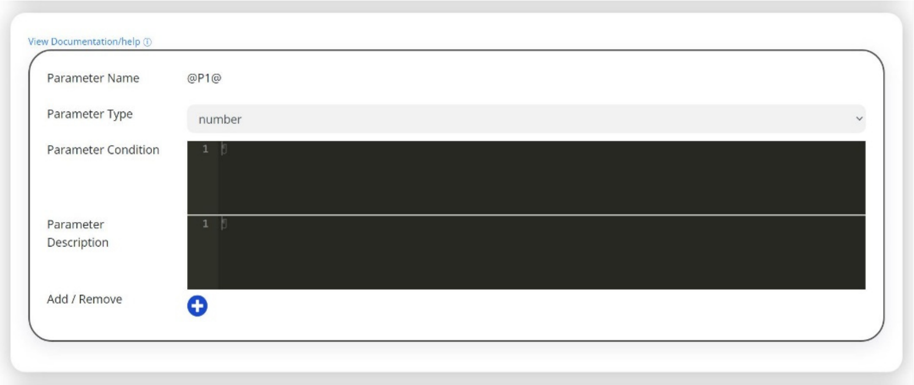
Type of output

Subject

Difficulty

Figure 3. General Information about the item.

The 'Parameter Name' is automatically given in the order of @P1@, @P2@, @P3@... There are a total of six types of 'Parameter types': number, select, code, Latex, text, and array. The code type is the only use to develop the item instances used in this study. In detail, when an item writer chooses the code type for the parameter type, writer can put any code of which the syntax follows general R programming guidelines in condition field. As the variables related to the time estimation are starting hour, starting minute, and time increase sequence, we use three parameters to control these three variables. For example, we set the 'Parameter Condition' of the first parameter of the time estimation item model, @P1@, as "sample(seq(10, 120, 10), 1)", which means choose a number among the vector between 10 to 120 with interval 10.



View Documentation/help

Parameter Name @P1@

Parameter Type number

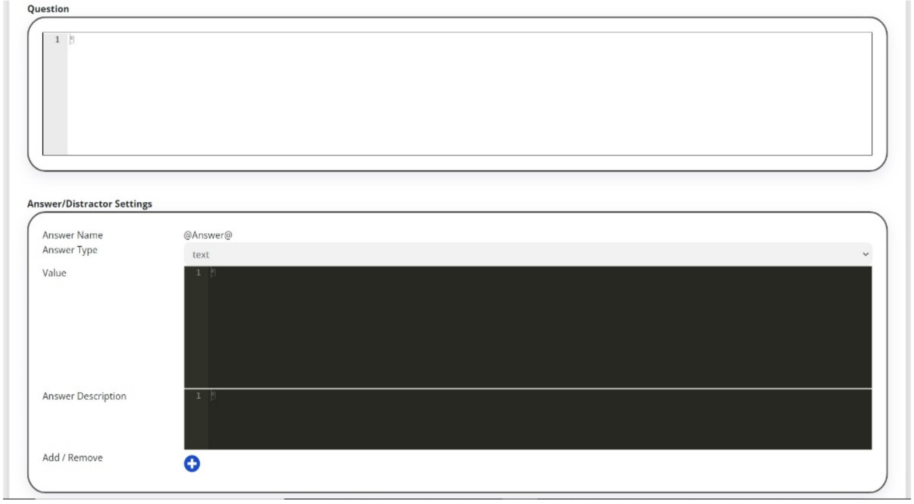
Parameter Condition

Parameter Description

Add / Remove

Figure 4. Parameterization.

The box in Figure 5 is a box for writing a question that will actually be presented to students. There is also a space where options (answer and distractors) are created, which can be added or removed through ‘Add/Remove’. There are six types of answer types, the same as the parameter types described above. When constructing options other than the correct answer, it is desirable for item generators to identify errors that may occur in the students’ cognitive process by considering the cognitive



Question

Answer/Distractor Settings

Answer Name @Answer@

Answer Type text

Value

Answer Description

Add / Remove

Figure 5. Question & Choices.

model. In the process of creating these options, the reason why the option is a correct answer (or not a correct answer) can be described in the ‘Answer Description’ to automatically present the reason for the wrong answer to students. This function is especially useful when a teacher/item writer shares it with other teachers/item writers.

What makes CLASS unique is the feature that allows item generators to use their own R code or Latex by using ‘User-defined code’ and ‘User-defined Latex’ spaces so that complex formulas or graphics can be worked separately from the item generation space (Figure 5). Item writers can create complex R codes in the left column of Figure 6 and Latex in the right column, which can be recalled as @CC@ and @CL@ variables, respectively, in the ‘Question’ section. For instance, we use the user-defined code to generate the item model instances of the time estimation and the paper-folding, especially to construct clocks and shapes. Also, if item writers write ‘\(\int_{@P2@}^{@P3@} x^@P1@ dx\)’ on the user-defined Latex space and type @CL@ in the question box presented in Figure 5, examinees can see an expression that means “dividing function x^(@P1@) by the interval from @P2@ to @P3@” when actually solving the question.

In addition, CLASS allows students to check immediate feedback and the process of solving the problem by writing them with the item model in advance. Since the

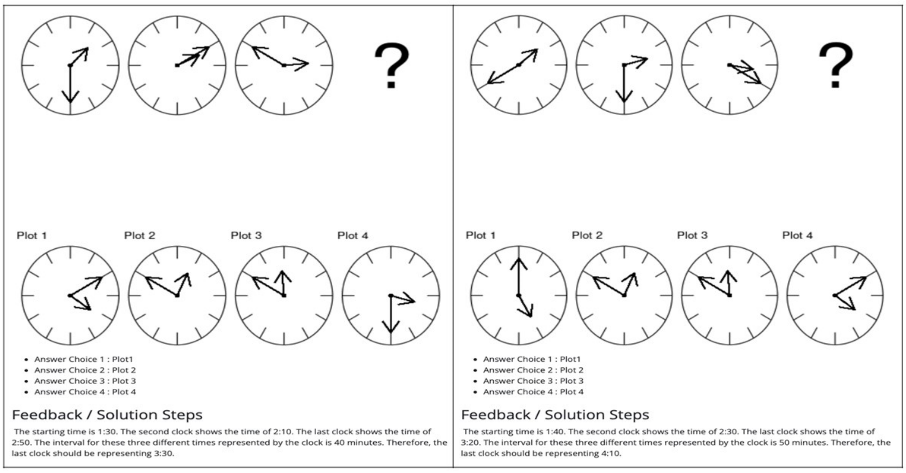
Figure 6. User-defined code and user-defined Latex.

parameters can be used in the feedback and problem-solving process, it is possible to write feedback in units of item models, in that it is convenient not to write feedback in units of each item. For example, we write ‘Feedback / Solution Steps’ as “The starting time is @P2@:@P3@. The second clock shows the time of @P5@:@P6@. The last clock shows the time of @P7@:@P8@. The interval for these three different times represented by the clock is @P4@ minutes. Therefore, the last clock should be representing @P9@:@P10@.”

Development Result

The concept of AIG is utilized to create many item instances from a single item model in CLASS. A single item model can create multiple instances of an item so that although the conditions are different, the concept of the item model remains the same for each item. The AIG in CLASS utilizes a computerized engine to create the different parameters and conditions that an item can be initialized with. All of these parameters and conditions are preset by item writers so that the flexibility and sample space can all be controlled by the item writer. The main difference in traditional item writing (based on paper-pencil testing design) and item writing with AIG (based on ECD and AE) is that traditional item writing has the sample space of the condition to be held constant with one measurement while AIG creates the flexibility for the engine to create multiple instances using this flexible sample space by utilizing variables.

We can visualize this concept by looking at one conceptual item generated by CLASS as shown in Figure 7. This item model below is indirectly asking the test taker what the subsequent time would be given that there is a pattern associated from the clocks above, with the question mark asking the test taker to correctly identify the next sequence of time. In this item model there are three main parameters that can be utilized for AIG. The three parameters are starting hour, starting minute, and time increase sequence. In the nature of the question, AIG has the ability to make an infinite number of item instances due to one of the parameters, time increase sequence, being an infinite number. Item writer has the ability to increase or decrease the sample space for items that can be generated by restricting any of the three parameters. In Figure 7, we can see that item instance 1 has 1, 30, and 40 for the parameters while item instance 2 has 1, 40, and 50 for the parameters. The writer of the clock item instance restricted the parameters with the following restrictions. The starting hour being a value between 0 and 2, the starting minute to be between 0 and 60, and the sequence parameter to be a number between 40 and 100, which results in a possible 11,163 item generations possible with AIG.



- Answer Choice 1 : Plot1
- Answer Choice 2 : Plot 2
- Answer Choice 3 : Plot 3
- Answer Choice 4 : Plot 4

Feedback / Solution Steps

The starting time is 1:30. The second clock shows the time of 2:10. The last clock shows the time of 2:50. The interval for these three different times represented by the clock is 40 minutes. Therefore, the last clock should be representing 3:30.

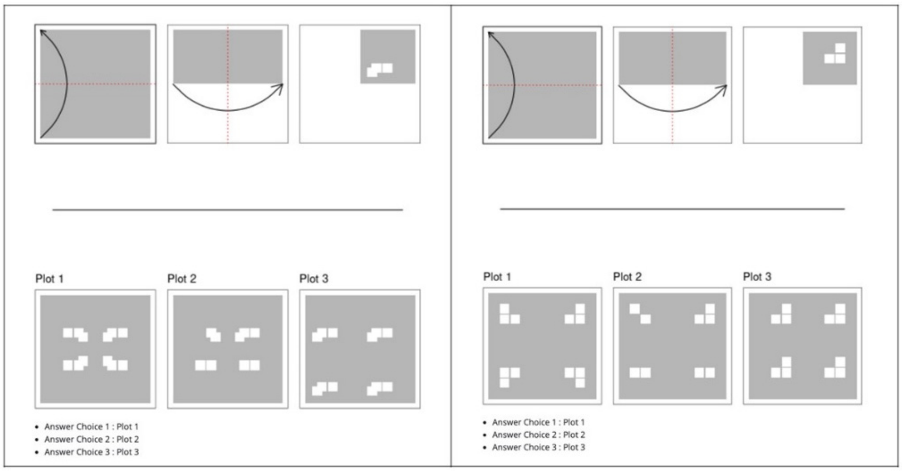
- Answer Choice 1 : Plot1
- Answer Choice 2 : Plot 2
- Answer Choice 3 : Plot 3
- Answer Choice 4 : Plot 4

Feedback / Solution Steps

The starting time is 1:40. The second clock shows the time of 2:30. The last clock shows the time of 3:20. The interval for these three different times represented by the clock is 50 minutes. Therefore, the last clock should be representing 4:10.

Figure 7. Time estimation item instances.

The possibilities of AIG don’t have to be in the realm of one-dimensional item parameter selection. Test makers can utilize AIG to make cognition test examples to fit their assessment needs. Figure 8 shows an item generated by AIG with many more parameters that can be flexibly managed. This item model shows what the final diagram will result in if a piece of paper was folded horizontally in half, vertically in half, and stamped. For this particular item model, the stamp location was the parameters that were given a sample size for AIG to initialize different item instances. The parameters were the coordinates of where three stamps were to be placed. Considering this is a plane with the dimensions of 1cm x 1cm the stamps being placed could be endless, which is why AIG can initialize different stamp locations to output various instances of the problem. However, if item writer wanted to increase the flexibility and sample space of this particular problem the item writer could even expand the parameters to which directions the folding occurs such as a folding direction of diagonal or along a certain line.



- Answer Choice 1 : Plot 1
- Answer Choice 2 : Plot 2
- Answer Choice 3 : Plot 3

- Answer Choice 1 : Plot 1
- Answer Choice 2 : Plot 2
- Answer Choice 3 : Plot 3

Figure 8. Paper-folding item instances.

Discussion

AIG revolutionized item generating practice and test design through item models. Item generation based on the traditional paper-pencil testing design has evolved into AIG based on new test design approaches, ECD and AE. The advantages of AIG can be divided into quantitative and qualitative aspects. In terms of quantity, AIG can generate a vast number of items in a short time, and in terms of quality, isomorphic items can be created with one item model to ensure the validity of the items and to reduce the errors. Developing a test through AIG gives the effect of preventing cheating in the test through the quantitative aspect advantage of it, in addition, through the qualitative aspect, problem solving or practice using memory is minimized to reduce errors in measurement. Also, because of the fact that item models are constructed based on cognitive models, items generated by those models get reliabilities and can be composed of meaningful distractors that identify examinee's misconception. Since cognitive ability is used as an indicator of judgment in fields such as education and medicine, the development of measures of cognitive ability using AIG greatly contributes to the development of cognitive test sites in various fields.

One of the theorical backgrounds of AIG is ECD which consist of five layers, including conceptual assessment framework. The three models, student model, evidence model, and task model, in conceptual assessment framework have built the basis of the process of AIG within CLASS. Although the three models of AE, construct map, task model, and template, are expressed differently with the three models of conceptual assessment framework, the implied meanings of each model are corresponding perfectly.

We choose the items which belong to CogAT, the renowned cognitive measurement test in United States. It is obvious that the validity of the items in the current study has been verified because these are items regarding CogAT, which has already been validated and widely used (Lakin, 2018) and these are selected in the field of CHC theory. This study showed the process of developing cognitive items in CogAT by using AIG in CLASS. In terms of the classification of the CHC theory, the two item instances selected in this study are the one belonging to quantitative reasoning in fluid reasoning and the other belonging to visualization in visual processing. For item instances of quantitative reasoning, an item model of estimating time was selected, and for item instances of visualization, a paper-folding item model was selected.

The first step to generate item models is that item writers should define the knowledge, skills, and abilities required to solve the task. In this study, these are quantitative reasoning and visualization. The second step is determining how we are going to measure these concepts. We choose time estimation and paper-folding instances, respectively. With determining those things, item writers should also construct cognitive models which represent examinees' processes of thinking in the second step to assure the reliabilities in order to determine the effectiveness and difficulty of items before generating items, to define parameters of item models having influence on variables we want to measure, and to make attractive distractors. The last step is visualizing the step 1 and 2. We use CLASS flatform to visualize for this third step.

In this paper, we demonstrated small numbers of cognitive item models by CLASS. However, it is scalable to provide more item models corresponding to all cognitive areas. As a future study, we plan to develop a cognitive test expanding areas of CogAT within the CHC theory to bolster research on cognition. Thus, we want to present recommendations for subsequent research on measures of cognitive ability using AIG. Using AIG, we may generate various cognitive ability measurement items in addition to the two item instances introduced above. Two additional cognitive items presented in Figure 9, belonging to inductive reasoning and quantitative reasoning among fluid reasoning, respectively. Although multiple items within CogAT are difficult to be developed in traditional testing design, isomorphic item instances can be easily produced by item models using AIG within CLASS. Using CLASS, items in other cognitive areas other than those presented in Figure 8 can be sufficiently developed. Additionally, the usability of AIG isn't limited to only cognitive domains, but it can be applied to any other fields, including math, statistics, or even reading and language arts.

Also, Since AIG within CLASS has been already used to produce items in the subjects of mathematics and statistics in the college education field, the availability of it can be widely expanded. So, verifying the availability by doing research with real

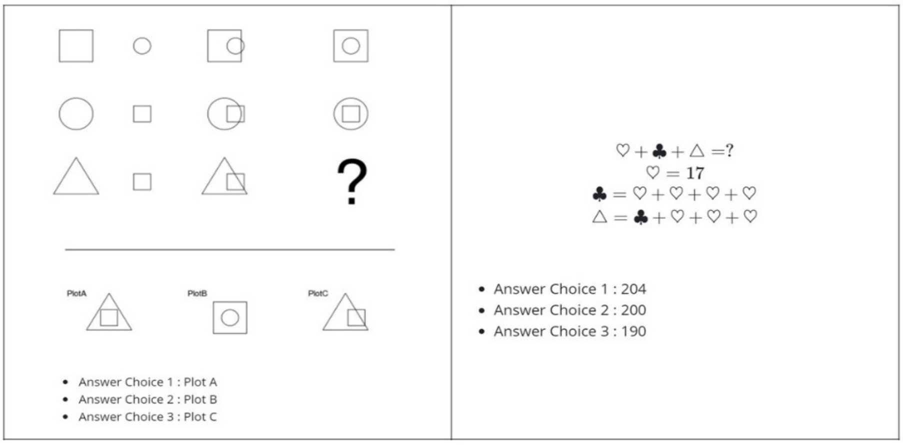


Figure 9. Other item instances.

examinees should be needed.

However, there are still several limitations of AIG. First, AIG requires more efforts to generate an item model but takes less after all. The more efforts mainly imply that AIG requires coding skills in addition to all skills that are used for traditional item writing. For supporting this limitation, in CLASS, it affords the guidelines and examples useful to build item models. Second, if an item writer makes wrong cognitive models or item models, the result, items, may not work well. However, Girel and Haladyna (2012) has already indicated that it is time-consuming that all new items must be field tested prior to operational use so that their psychometric properties can be documented and many of those do not perform as intended and, therefore, must be either revised or discarded. In this respect, AIG is the better option for developing new items. Moreover, to prevent constructing inaccurate item models, CLASS offers the function which warn errors in item model codes. Also, it has the expert approval procedure to register item models on the common questions list. Finally, to justify the reliability and validity of items generated by AIG in the current study, additional verification study should be done further. It would be helpful to check the verification study of other scale scores for cognitive domains already done (Ryoo et al., 2022).

In summary, by presenting the process of the procedure of AIG in cognitive field, we could check the possibility of applying the new, innovative test development method in real life. In term of academic achievement or development and clinical diagnosis, the cognitive measurement is a crucial indicator, thus the reliable item generation process based on item models is the key of making accurate cognitive measurement tests.

References

Anderson, L. W., & Krathwohl, D. R. (2001). A taxonomy for learning, teaching, and assessing: *A revision of Bloom's taxonomy of educational objectives.* Longman.

Behrens, J. T., Mislevy, R. J., DiCerbo, K. E., & Levy, R. (2010). An evidence centered design for learning and assessment in the digital world. CRESST Report 778. *National Center for Research on Evaluation, Standards, and Student Testing* (CRESST).

Bejar, I. I. (2002). Generative testing: From conception to implementation. In S. H. Irvine & P. C. Kyllonen (Eds.), *Item generation for test development* (pp.199-217). Erlbaum.

Bloom, B. (1956). Bloom's taxonomy.

Bryan, V. M., & Mayer, J. D. (2020). A meta-analysis of the correlations among broad intelligences: Understanding their relations. *Intelligence, 81.*

Carmines, E. G., & Zeller, R. A. 1979. *Reliability and Validity Assessment.* SAGE.

CLASS [website]. (2022.01.28.). URL: https://class-analytics.com/

Drasgow, F., Luecht, R. M., & Bennett, R. (2006). Technology and testing. In *Educational measurement* (4th ed.). American Council on Education/Praeger Publishers.

Embretson, S. E. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological methods, 3*(3), 380.

Embretson, S., & Yang, X. (2006). 23 Automatic item generation and cognitive psychology. *Handbook of statistics*, *26*, 747-768.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). John Wiley & Sons, Inc.

Gierl, M. J., & Haladyna, T. M. (Eds.). (2012). *Automatic item generation: Theory and practice.* Routledge.

Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced methods in automatic item generation* (1st ed.). Routledge.

Goldberg, T. E., Harvey, P. D., Wesnes, K. A., Snyder, P. J., & Schneider, L. S. (2015). Practice effects due to serial cognitive assessment: implications for preclinical Alzheimer's disease randomized controlled trials. *Alzheimer's* & *Dementia: Diagnosis, Assessment* & *Disease Monitoring, 1*(1), 103-111.

Irvine, Sidney H., & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development* (1st ed). Routledge.

Jutten, R. J., Rentz, D. M., Fu, J. F., Mayblyum, D. V., Amariglio, R. E., Buckley, R. F., Properzi, M. J., Maruff, P., Stark, C. E., Yassa, M. A., Johnson, K. A., Sperling, R. A., & Papp, K. V. (2021). Monthly at-home computerized cognitive testing to detect diminished practice effects in preclinical Alzheimer's disease. *Frontiers in aging neuroscience, 13*.

Lakin, J. M. (2018). Making the Cut in Gifted Selection: Score Combination Rules and Their Impact on Program Diversity. *Gifted Child Quarterly, 62*(2), 210-219. https://doi.org/10.1177/0016986217752099

Latex [website]. (2022.02.10.). URL: https://www.latex-project.org/

McGrew, K. S. (2009). CHC theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*, *37*(1), 1-10.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2004). A brief introduction to evidence-centered design. CSE Report 632. *US Department of Education.*

Mislevy, R. J., Steinberg, L. S., & Almond, R. G. (2003). Focus article: On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*(1), 3-62.

Osborne, C. (1991). Statistical calibration: A review. *International Statistical Review /Revue Internationale de Statistique*, 309-336.

Park, H. J., Ryoo, H. S., Kwon, J., & Ryoo, J. H. (2022). Change of paradigm on LMS for online education: LMS implementing learning analytics and online assessment. *The Educational Research for Tomorrow*, *35*(2), 49-72.

Riverside Insights [website]. (2022.02.05). URL:

https://www.riversideinsights.com/home

Ryoo, J. H., Park, S., Suh, H., Choi, J., & Kwon, J. (2022). Development of a new measure of cognitive ability using automatic item generation and its psychometric properties. *SAGE Open*, *12*(2), 21582440221095016.

Schneider, W. J., & McGrew, K. S. (2012). The Cattell-Horn-Carroll model of intelligence. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Test, and Issues* (pp. 99-144). Guilford Publications.

Schneider, W. J., & McGrew, K. S. (2018). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary Intellectual Assessment. Theories, Tests, and Issues* (4th ed., pp. 73-163). The Guilford Press.

Stanek, K. M., Gunstad, J., Spitznagel, M. B., Waechter, D., Hughes, J. W., Luyster, F., ... & Rosneck, J. (2011). Improvements in cognitive function following cardiac rehabilitation for older adults with cardiovascular disease. *International Journal of Neuroscience*, *121*(2), 86-93.

Taherdoost, H. (2016). Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. *How to test the validation of a questionnaire/survey in a research (August 10, 2016).*

Thompson, B. (2011). What is the CogAT(Cognitive Abilities Test) and Why use it?, *Homeschool Handbook*.

Warnimont, C. (2010). *The relationship between students' performance on the cognitive abilities test (COGAT) and the fourth and fifth grade reading and math achievement tests in Ohio* (Unpublished Doctoral dissertation). Bowling Green State University, USA.

Wesnes, K., & Pincock, C. (2002). Practice effects on cognitive tasks: A major problem?. *The Lancet Neurology, 1*(8), 473.

Seungyeon Sa

Graduate Student, Dept. of Education, College of Educational Sciences, Yonsei University. Interests: Educational Statistics/Data Science, Longitudinal Data Analysis, Learning Management System, Automatic Item Generation

E-mail: lucky2021@yonsei.ac.kr

Hyun Suk Ryoo

Graduate Student, School of Data Science, University of Virginia. Interests: Machine Learning, Data Mining, and Quantitative Methods

E-mail: hr2ee@virginia.edu

Homepage: Hyun Suk Ryoo (hyunsuk-ryoo.com)

Ji Hoon Ryoo

Associate Professor, Dept. of Education, College of Educational Sciences, Yonsei University. Interests: Educational Statistics/Data Science, Learning Analytics, Learning Management System

E-mail: ryoox001@yonsei.ac.kr

Homepage: https://ryoox001.github.io/

Received: September 10, 2022 / Peer review completed: September 26, 2022/ Accepted: October 10, 2022