MODESTUM EURASIA Journal of Mathematics, Science and Technology Education, 2023, 19(6), em2282
ISSN:1305-8223 (online)
OPEN ACCESS Research Paper https://doi.org/10.29333/ejmste/13251

Development of the automatic item generation system for the diagnosis of misconceptions about force and laws of motion

Kittitas Wancham [1], Kamonwan Tangdhanakanond [1*], Sirichai Kanjanawasee [1]

[1] Faculty of Education, Chulalongkorn University, Bangkok, THAILAND

Received 14 March 2023 ▪ Accepted 27 April 2023

Abstract

The understanding of force and laws of motion is a fundamental foundation for learning mechanics and understanding other complex physics-related subjects. Automatic item generation (AIG) is also suitable for generating items and able to reduce the chance of item exposure. We, thus, developed an AIG system for the diagnosis of misconceptions about force and laws of motion in order to create a large number of quality items that would be used to diagnose students' misconceptions. AIG system that has been developed contains 18 item models; it can generate 320-3,200 test items. The system contains six menus, i.e., (1) users' data, (2) item models, (3) item generation, (4) test generation, (5) the users' guide, and (6) the system's developer. Based on the examination of AIG system's quality by experts on educational assessment and experts on information technology, AIG's quality in terms of utility, feasibility, propriety, and accuracy is at the highest level. The system was improved using the two dimensions of users' experiences with physics instructors, i.e., (1) pragmatic dimension and (2) hedonic dimension. This research offers an approach to developing AIG system that responds to users' needs.

Keywords: automatic item generation, item model, force and laws of motion, misconception

INTRODUCTION

The concept of force and Newton's laws of motion is fundamental to learning mechanics and other complicated concepts in physics (Saglam-Arslan & Devecioglu, 2010). This concept is also one of the core concepts in STEM education that students should have deep understanding on without holding misconceptions, so that they can learn in this field effectively (Pellegrino & Hilton 2012; Thibaut et al., 2018). Additionally, the force and Newton's laws of motion plays an essential role in elucidating physical phenomena that we experience in everyday life. That is, it depicts a relationship between an object's motion and forces acting on it (Sornkhatha & Srisawasdi, 2013). If students have misconceptions about force and Newton's laws of motion, they will not succeed in learning physics (Aini et al., 2021).

Wancham et al. (2023) developed the diagnostic test for misconceptions about force and laws of motion by applying the cognitively diagnostic assessment (CDA). CDA is a type of educational assessment designed to diagnose predestined set of attributes to provide detailed diagnostic information to individual students about their strengths and weaknesses and to provide useful information to teachers that will help them design remedial programs (de la Torre & Minchen, 2014; Javidanmehr & Sarab, 2017). This diagnostic test has good psychometric properties. It consists of six attributes, which are

(1) resultant force,
(2) Newton's first law of motion,
(3) Newton's second law of motion,
(4) Newton's third law of motion,
(5) frictional force, and
(6) the gravitational force.

Although the diagnostic test provides useful information for designing remedial programs to correct misconceptions about force and laws of motion, using the diagnostic test for multiple administrations results in item exposure (Gierl et al., 2008). To overcome this problem, a testing system should have a large item bank that provides test items for each administration.

© 2023 by the authors; licensee Modestum. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/).
✉ wkittitas@outlook.com ✉ tkamonwan@hotmail.com **(*Correspondence)** ✉ sirichaik@chula.ac.th

Contribution to the literature

- AIG system for the diagnosis of force and laws of motion helps generate a large number of quality parallel items to diagnose students' misconceptions. The results will be used to correct students' misconceptions about force and laws of motion.
- Collecting data on users' experiences helps improve AIG system that responds to users' needs.
- AIG system should be available for access through various browsers and adjustable depending on the device used. The menu bars must be arranged according to the system's operational sequence so that it's user friendly. The system should contain only necessary menu bars that are arranged in a clean, organized manner. The icons for the menu bars should be indicative of their functions. Users must be able to revise the item models and all the necessary details, depending on users' varied needs.

Moreover, the availability of a large item bank enhances the security of the testing system and fairness in testing. However, traditional test development cannot satisfy the demand for a large item bank, as the traditional approach is time-consuming and costly. That is, each item is written, reviewed, revised, piloted, and evaluated in terms of psychometric properties. An approach used to fulfil this demand is automatic item generation (AIG) (Embretson & Yang, 2006; Gierl & Lai, 2013).

AIG is a suitable device to generate test items as it can generate many items from the item model, which makes the cost of generating test items lower than that of generating each test item. The system can quickly create a high-quality item bank and reduce the chance of item exposure due to the large item bank. In addition, AIG also reduces the burden for the test creator, simplifies the process of item review by experts, and reduces the amount of the item pilots (Gierl & Lai, 2016; Gierl et al., 2008; Graf et al., 2005; Sinharay & Johnson, 2013). We, thus, developed AIG system to diagnose misconceptions about force and laws of motion to create a large number of high-quality items to diagnose students' misconceptions, which will be used as data for planning remedial programs and adjust students' misconceptions about force and laws of motion.

LITERATURE REVIEW

This section provides details about misconceptions about force and laws of motion and AIG that comprises four topics, i.e., misconceptions about force and laws of motion, definition of AIG, item model, and steps of AIG process. These details are presented as follows.

Misconceptions About Force and Laws of Motion

Misconceptions refer to ideas, beliefs, and understandings that contradict scientific concepts (Kaniawati et al., 2019; Narjaikaew, 2013). Scientific misconceptions can be divided into five categories, i.e,

(1) preconceived notions,

(2) nonscientific beliefs,

(3) conceptual misunderstanding,

(4) factual misconceptions, and

(5) vernacular misconceptions.

Preconceived notions are notions that are based on experiences in daily life. Preconceived notions affect students' perspectives on scientific conceptions. Nonscientific beliefs are beliefs students acquire from other sources in addition to scientific education, such as religion.

Conceptual misunderstandings are understandings of scientific theories that contradict scientific explanations, which occur when students receive scientific education that does not encourage confrontation with contradictions, as a result of students' preconceived notions and nonscientific beliefs.

Factual misconceptions are misconceptions that result from learning about incorrect facts in their childhood that students still hold on to in their adulthood. Vernacular misconceptions are misconceptions that result from using words whose meanings differ when used in daily life and in scientific contexts (National Research Council, 1997; Yasri, 2014).

Misconceptions in students will increasingly hinder their learning of conceptions, as they function as the basis for their learning. Students will not be able to link up their existing knowledge with new knowledge, which leads to failure in learning. Students will also wrongly apply misconceptions into their daily life, in addition to losing their motivation to learn (Gurel et al., 2015).

Wancham et al. (2022) synthesized and grouped common misconceptions about force and laws of motion possessed by students in secondary and higher education. They categorized 27 misconceptions into six categories based on the force and laws of motion topics, which were

(1) resultant force,

(2) Newton's first law of motion,

(3) Newton's second law of motion,

(4) Newton's third law of motion,

(5) frictional force, and

(6) gravitational force.

Table 1 shows misconceptions in each category.

Table 1. Misconceptions about force & laws of motion

Category	Misconception
1. Resultant force	1.1 An object moves in the direction of the greater force. 1.2 An object changes its direction in the direction of the last force.
2. Newton's first law of motion	2.1 An object stores an applied force into an impetus to keep object going after the force is worn out. 2.2 An impetus keeps objects moving. 2.3 A trajectory of an object depends on an impressed impetus.
3. Newton's second law of motion	3.1 If there is no motion, there is no force acting on an object. 3.2 A moving object stops when the force is stopped. 3.3 If there is motion, there is a force acting on an object in its direction of motion. 3.4 If there is a force acting on an object at rest, the object will move. 3.5 When an object is moving, there is a force in the direction of its motion. 3.6 There is a linear relationship between force and velocity. In other words, a constant velocity results from a constant force. 3.7 An object that moves with a constant acceleration requires a constantly changing force. 3.8 Forces are caused by living or moving things. 3.9 Forces can only be caused by something touching an object.
4. Newton's third law of motion	4.1 An action-reaction pair of force acts on the same object. 4.2 According to applied forces between two objects, the greater mass exerts the greater force. 4.3 According to applied forces between two objects, the bigger object exerts the greater force. 4.4 According to applied forces between two objects, the most active object exerts the greater force. 4.5 When an object moves into an obstacle, the obstacle redirects or stops motion, but it cannot be the agent of an applied force.
5. Frictional force	5.1 Frictional force acts on an object when it moves. 5.2 Frictional force always acts opposite to the direction of motion. 5.3 Static frictional force is minimum when an object begins to move. 5.4 Static frictional force is constant & equals a coefficient of static friction multiplied by a normal force.
6. Gravitational force	6.1 For free fall, a heavier weight causes a bigger acceleration. In other words, heavier objects fall faster. 6.2 There is the gravitational force acting on an object when it is only on the earth. 6.3 The gravitational force has constant value and is the same everywhere. 6.4 The gravitational force does not act until an impetus wears down.

Definition of Automatic Item Generation

AIG is a process that integrates the cognitive theory and psychometrics and uses the item model to create test items by means of computer technology that collects all the probable elements determined in the item model. The process results in a quick generation of a large number of meaningful test items. The items can be generated in real time, in accordance to demand or as students are taking the test on the fly (Embretson & Yang, 2006; Gierl & Lai, 2016; Lai et al., 2016).

Item Model

The item model is the template that identifies the item attributes used to create items with equivalence. It is divided into three parts, i.e., stem, option, and auxiliary information. For stem and option, there are elements that are either strings or integers, both of which are variables in the item model that is used to create a large number of new items (Gierl & Lai, 2013; Graf et al., 2005; Sinharay & Johnson, 2008). Each item model contains details (Gierl & Lai, 2018; Gierl et al., 2008), as follows:

Stem is part that identifies the situation, content, and question to which students need to provide an answer.

Option is the part that determines the details of distractors and key.

Auxiliary information collects additional information about both the stem and option, which are required to create items. It consists of texts, pictures, tables, diagrams, sounds, and videos.

Constructed-response item model only creates stem. With multiple-choice item model, both stem and option will be created. Auxiliary information is not required.

The item model is divided into two types in accordance with information in the stem, i.e., the 1-layer item model and the n-layer item model. The details (Gierl & Lai, 2013; Lai et al., 2016) are, as follows:

The 1-layer item model is the item model that contains elements in the one layer and generates a linear item. The goal of the item generation using the 1-layer item model is to generate an item that interacts with a small number of elements in the item model. The items generated contains similar psychometric properties and can predict the item's psychometric properties based on the elements in the item model. The limitation is that the items generated will be too similar to each other, as they contain a small number of elements during the interaction. The items are, thus, called clones, ghost items, Franken-items, or siblings. The items considered to be clones are suitable for generating parallel tests and competency-based education for testing students multiple times with the same content using similar tests.

The n-layer item model is the item model with elements in the stem with over two levels. One element is nested in another element, which makes the item

generation a non-linear process. Item generation using the n-layer item model is better at generating items than the 1-layer item model. The goal of generating items using the n-layer item model is to generate items that put a large number of elements in the item model, which results in varied items, even though they are problematic when it comes to predicting the item's psychometric properties. A traditional approach to test development is, thus, necessary.

Steps of Automatic Item Generation Process

The process of AIG is divided into three steps, i.e.,

(1) determining the content for the item model,

(2) generating the item model, and

(3) generating items and evaluating the similarity of the items.

The steps (Gierl & Lai, 2013, 2016; Graf et al., 2005) are detailed as follows:

Determining the content for the item model

The first step of AIG is determining the content for the item model, which considers the design approach, knowledge, experience, theories, and research that demonstrates knowledge or skills students need to answer the items. The determination of the content for the item model requires the cognitive model as the framework for the item generation. The cognitive model provides details about knowledge and skills students need to answer the items, as well as the content that affect the difficulty of the items that result in the generation of the new item. Students' test-taking skill, thus, correlates with the interpretation of students' scores. Cognitive model is divided into three parts, i.e.,

(1) the relevant problems and situation or misconceptions that needed to be assessed,

(2) the sources of information–which can be those that is relevant to the problems or the general sources of information that can be applied to other problems–that are in line with the problems, and

(3) the features of the information, which comprise elements that need manipulation and the constraint of elements to make the content of the items meaningful.

All three components are obtained from the quality analysis of the parent item. The parent item is prototype for the item model generation.

Generating the item model

The generator of the item model must use the information obtained from the content determination process as an approach for the item model generation to create a new item. After creating an item model, the next step is an assessment of the content and feasibility specified in the cognitive model by the experts, as well as the structure of the item in the item model. Then, the cognitive model and the item model will be revised in accordance with experts' suggestions. The revision of the cognitive model and the item model is intended to improve the structure of the item model to ensure it is proper and feasible, so that it can generate quality and reasonable items. Gierl and Lai (2016) proposed that three aspects of the cognitive model and item model that should be assessed are content, logic, and presentation.

Generating items and evaluating similarity of items

This step relies on a computer program to generate items, using an item model that contains all probable elements specified in the cognitive model to generate a large number of meaningful items in a short period of time. Then, the similarity of the items generated from the same item model will be assessed using CSI (cosine similarity index). CSI ranges from 0 to 1, with 0 referring to both test items having no common words and 1 referring to both items containing similar words. If the item's CSI has a high average and low standard deviation (SD), then the item model will generate items that are isomorphs or clones. That is, items generated from the item model contains multiple repetitive words. If the item's CSI has a low average and high SD, then the item model will generate items that are variants, meaning items generated from the item model contains few repetitive words.

METHOD

Informants

The informants are divided into two groups, i.e.,

(1) the informants for the assessment of the quality of the cognitive model and the item model, which consists of seven experts in physics teaching and

(2) the informants for the assessment of the quality of the system and the user guide for AIG system.

The assessment requires four experts in educational assessment and three information technology experts.

Participants

20 high-school physics teachers with an average age of 31.30 (SD=4.18), 10 of whom are male and another ten females. Participants were selected based on purposive sampling. The sample size was determined according to Tang et al. (2021), who stated that the study on users' experiences needs at least 15 individuals as samples.

Materials

The research materials consist of

(1) the cognitive model assessment form for AIG and the item model,

(2) the assessment form for AIG system, and

(3) the assessment form for the user guide of AIG system.

The details, are as follows:

Cognitive model assessment form for automatic item generation and item model

We revised scoring rubric developed by Gierl and Lai (2016). There are three aspects for the assessment, i.e.,

(1) content (the details in the model suitable for assessing the attributes needed to be assessed),

(2) logic, (the combination of the content in the model to ensure the knowledge and skills are measured correctly), and

(3) presentation (the linguistic and grammatical verity of the item created).

There are four score levels, i.e., 1 is *not accepted*; 2 is *not accepted with major revision*; 3 means *accepted with minor revision*; and 4 means *accepted*. The cognitive model for AIG and the item model that are of quality must achieve level 3 scores in each aspect.

Assessment form for automatic item generation system

The assessment of AIG system consists of 14 question items with a five-rating scale, with 1 referring to *the lowest*, 2 *low*, 3 *average*, 4 *high*, and 5 *the highest*. There are four assessment dimensions, i.e.,

(1) utility (AIG system that responds to users' needs)

(2) feasibility (AIG system is applicable to real-world situations, convenient to use, and worth using),

(3) propriety (AIG system is in line with the process of AIG and does not affect stakeholders in a negative way), and

(4) accuracy (AIG system is able to generate items that are accurate and meaningful).

The results of the assessment's content validity showed that the test items contain the IOC ranging .86-1.00 (CVI=.93).

Assessment form for the user guide of automatic item generation system

The assessment form for the user guide of AIG system consists of 15 question items, with a five-rating scale, with 1 referring to *the lowest*, 2 *low*, 3 *average*, 4 *high*, and 5 *the highest*. There are four assessment dimensions. They are

(1) utility (guide for using AIG system responds to users' needs),

(2) feasibility (guide for using AIG system is applicable to real-world situations, convenient to use, and worth using),

(3) propriety (guide for using AIG system is in line with the process of automatic test generation), and

(4) accuracy (guide for using AIG system is accurate in accordance with the item generation system).

The results of the assessment's content validity showed that the test items contain the IOC ranging .86-1.00 (CVI=.94).

Procedure

1. Experts in physics teaching were required to assess the quality of the cognitive model and the item model using the cognitive model assessment form for AIG and the item model. The cognitive model and item model were revised according to the experts' suggestions.
2. Experts in educational assessment and experts in information technology assessed the quality of the system and the guide for the usage of AIG system for the diagnosis of misconceptions about force and laws of motion. They used assessment form for AIG system and the assessment form for the user guide of AIG system. The system and guide were revised according to experts' suggestions.
3. AIG for the diagnosis of misconceptions about force and laws of motion and the guide for using the system were given to 20 high-school physics instructors as part of the collection of data on users' experiences, in both the pragmatic dimension–to consider the utility–and the hedonic dimension–to consider interests and impressions by interviewing teachers about their experiences with the system usage, problems in using the system, the system's elements they like, what needs to be revised and added to the system, and the content and presentation in the guide for using the system. This also includes what needs to be revised or added to the guide for the system usage. Then, the system and the guide for AIG system usage were revised accordingly to ensure they respond to users' needs.

Data Analysis

1. Analyze the quality of the cognitive model for AIG and the item model using frequency.
2. Analyze the similarity of the items in each item model using CSI index under the *spatialEco* package in R program and analyze the minimum, maximum, average, as well as SD of CSI index of each item model.
3. Analyze the quality of the system and the guide for using AIG system using frequency, arithmetic mean (M), and SD.
4. Analyze content based on interviews with high-school physics teachers on user-experience data and guide for using AIG using content analysis.

Table 2. Q-matrix for creating the cognitive model & the item model

Attribute	Item																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1.1	1	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	0
1.2	0	0	1	0	1	0	0	0	0	1	0	1	0	0	0	0	0	0
2.1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0
2.2	0	1	1	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0
2.3	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
3.2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
3.3	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
3.4	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3.5	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
3.6	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
3.7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
3.8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
3.9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
4.1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	1
4.2	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
4.3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
4.4	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
4.5	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	0	0
5.1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1
5.2	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0
5.3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
5.4	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
6.1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
6.2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
6.3	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
6.4	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Note. Each attribute contains sub-points as detailed in **Table 1**

RESULTS

Results of Generating & Analyzing Quality of Cognitive Model for Automatic Item Generation & Item Model

We created cognitive model for AIG and item model in accordance with Q-matrix developed by Wancham et al. (2023), as shown in **Table 2**. Q-matrix is a table that identifies relationship between attributes and items. The table contains number 1 and 0. That is, 1 represents items that assess attributes, while 0 represents items that do not assess any attributes. Based on **Table 2**, there are six attributes intended for assessment, which are

(1) resultant force,

(2) Newton's first law of motion,

(3) Newton's second law of motion,

(4) Newton's third law of motion,

(5) frictional force, and

(6) gravitational force.

There are 18 test items, which requires 18 model pairs of the cognitive model for AIG system and item model. The example of the cognitive model is shown in **Figure 1**, which is the cognitive model for model 4 of AIG.

The item model is shown in **Figure 2**, which is model 4 of the item model. The assessment results of all 18 model pairs of the cognitive model for AIG and the item model, with three aspects of assessment (content, logic, and presentation) showed that every pair of the cognitive model for AIG and the item model achieve level 3 scores (*accepted with minor revision*) in all aspects. Most of the model pairs achieved level 4 scores (*accepted*) in all aspects.

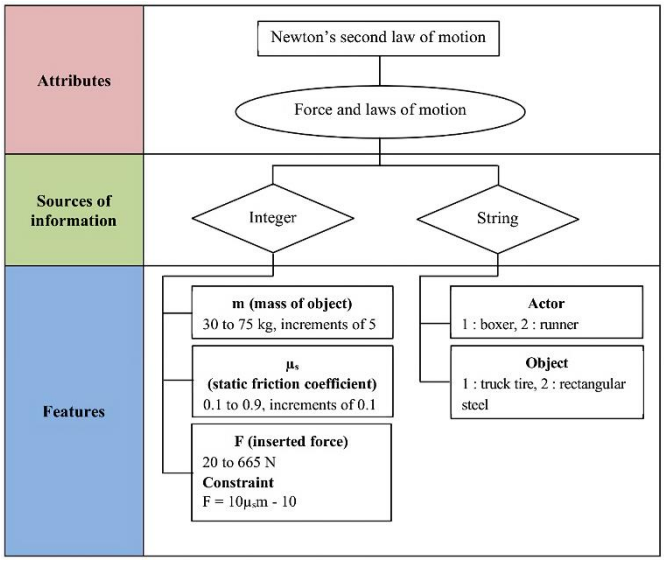

Figure 1. An example of the cognitive model (Source: Authors' own elaboration)

The 18 item models can generate between 320-3,200 test items (M=976.33, SD=918.61). In addition, based on

Parent item	A runner ties himself with a rope to a rectangular steel with a mass of 50 kilograms. The rectangular steel lays still on a flat ground with a static friction coefficient of 0.2. The runner, then, inserts a force of 90 newtons on the rectangular steel to the horizontal line. Please find out if the runner can tow the rectangular steel and give reasons to support the answer.
Item model	
Stem	A **[Actor]** ties himself with a rope to a **[Object]** with a mass of **[m]** kilograms. The **[Object]** lays still on a flat ground with a static friction coefficient of **[μ_s]**. The **[Actor]**, then, inserts a force of **[F]** newtons on the **[Object]** to the horizontal line. Please find out if the **[Actor]** can tow the **[Object]** and give reasons to support the answer.
Elements	**[Actor] (string)** : 1 : boxer, 2 : runner **[Object] (string)** : 1 : truck tire, 2 : rectangular steel **[m] (integer)** : 30 to 75, increments of 5 **[μ_s] (integer)** : 0.1 to 0.9, increments of 0.1 **[F] (integer)** : 10[μ_s][m] - 10
Key	The **[Actor]** cannot drag the **[Object]** because the inserted force is outweighed by the maximum static friction, which is equivalent to 10**[μ_s]**[m] newtons. The **[Object]** is, thus, not moved.

Figure 2. An example of the item model (Source: Authors' own elaboration)

Table 3. CSI & number of generated items of item models

Item model	CSI				Number of generated item
	M	SD	Min	Max	
1	0.78	0.11	0.62	1.00	3,200
2	0.61	0.21	0.38	0.97	600
3	0.63	0.20	0.38	1.00	3,200
4	0.72	0.15	0.52	0.97	360
5	0.62	0.19	0.42	1.00	720
6	0.62	0.16	0.44	0.99	320
7	0.78	0.14	0.62	1.00	800
8	0.78	0.15	0.61	0.98	350
9	0.78	0.14	0.59	1.00	640
10	0.65	0.28	0.33	1.00	1,600
11	0.75	0.18	0.55	0.99	1,920
12	0.63	0.18	0.41	1.00	864
13	0.62	0.16	0.47	0.98	340
14	0.73	0.22	0.47	1.00	320
15	0.78	0.13	0.65	0.99	600
16	0.80	0.12	0.61	0.98	360
17	0.72	0.12	0.61	0.99	480
18	0.79	0.10	0.68	0.99	900

CSI for assessing the similarity of the items generated from each item model, the item models have a CSI average of 0.61-0.80 and an SD of the CSI index ranging 0.10-0.20, as shown in **Table 3**, meaning the items generated from each item model are not too similar or different. The items generated from each item model are parallel items.

Results of Developing & Verifying Quality of Automatic Item Generation System for Diagnosis of Misconceptions About Force & Laws of Motion

Results in this part are divided into two aspects, i.e.,

(1) the details of AIG system for the diagnosis of misconceptions about force and laws of motion and

(2) the results of verifying the quality of AIG system for the diagnosis of force and laws of motion.

Each aspect contains details, as follows.

Details of automatic item generation system for the diagnosis of misconceptions about force and laws of motion

AIG system for the diagnosis of misconceptions about force and laws of motion that was revised in accordance with suggestions by experts on educational assessment, experts in information technology, and physics teachers are, as follows:

AIG item for the diagnosis of misconceptions about force and laws of motion operates as a database on an internet network developed using Laravel, which is a PHP web application framework with MySQL. AIG system can be accessed through browsers such as Chrome, Google Microsoft Edge, Mozilla Firefox, and Safari. The display can be optimized for different devices, such as computers, smartphones, and tablets. The function of AIG system can be divided into two parts, i.e., the usage for users and the usage for admins. Both parts function similarly, but the admins have an additional duty of managing users' data. The main components of AIG system contain details, as follows:

Usage for users: Using AIG system contains two primary parts, i.e., the login page and the main window for using AIG system. The login page contains three parts, i.e.,

(1) registration,

(2) login, and

(3) password retrieval.

While the main window for using AIG system contains six menu bars, which are

(1) users' data,

(2) item model,

(3) item generation,

(4) test generation,

(5) user guide, and

(6) system developer.

Each menu bar's function is detailed as follows:

Users' data: The user's data menu bar shows details filled out by the users in the registration form. There are eight items, which are

(1) users' account,

(2) a new password,

(3) the confirmation of a new password,

(4) the name and last name,

(5) academic institution/workplace,

(6) occupation,

(7) mobile phone number, and

(8) email.

The information can be changed by the user.

Item model: The item generation and the arrangement of test forms for the diagnosis of misconceptions about force and laws of motion involve three menu bars, i.e., item model, item generation, and test generation. The item model menu contains details of 18 item models for the diagnosis of misconceptions about force and laws of motion. Each model contains details in six aspects, as follows:

1. **The description:** The description header is used to record details about the item models, such as the attributes that need to be assessed.
2. **Parent item:** The parent item header is used to demonstrate items that are prototypes for the item model generation.
3. **Stem:** The stem header is used to show details about the situations and questions students need to find answers to.
4. **Elements:** The element header is used to demonstrate and modify the element values. There are three attributes of the elements, i.e.,

   (1) elements that are texts, which are shown in the green tab,

   (2) elements that are alphabets, which are shown in the blue tab, and

   (3) elements that are numbers, which are shown in the red tab.

   Users can adjust the settings of the elements that are alphabets and numbers only.
5. **Key:** The key header is used to determine details about answers for each item in the item model.
6. **Pictures:** The picture header is used to determine the details of the photo display of each item in the item model. The picture header is included only in some models.

Examples of the item model menu bar, as shown in **Figure 3**.

All 18 item models contain complete details. That is, users can generate test items without having to revise any details in the item model. However, users can revise four sections of the information before the item is generated. They are

(1) information in the description header,

(2) information about the elements,

(3) information about the keys, and

(4) information about the photos.

If users want to revise any elements, they have to click on the element bar they want to fix, then a new

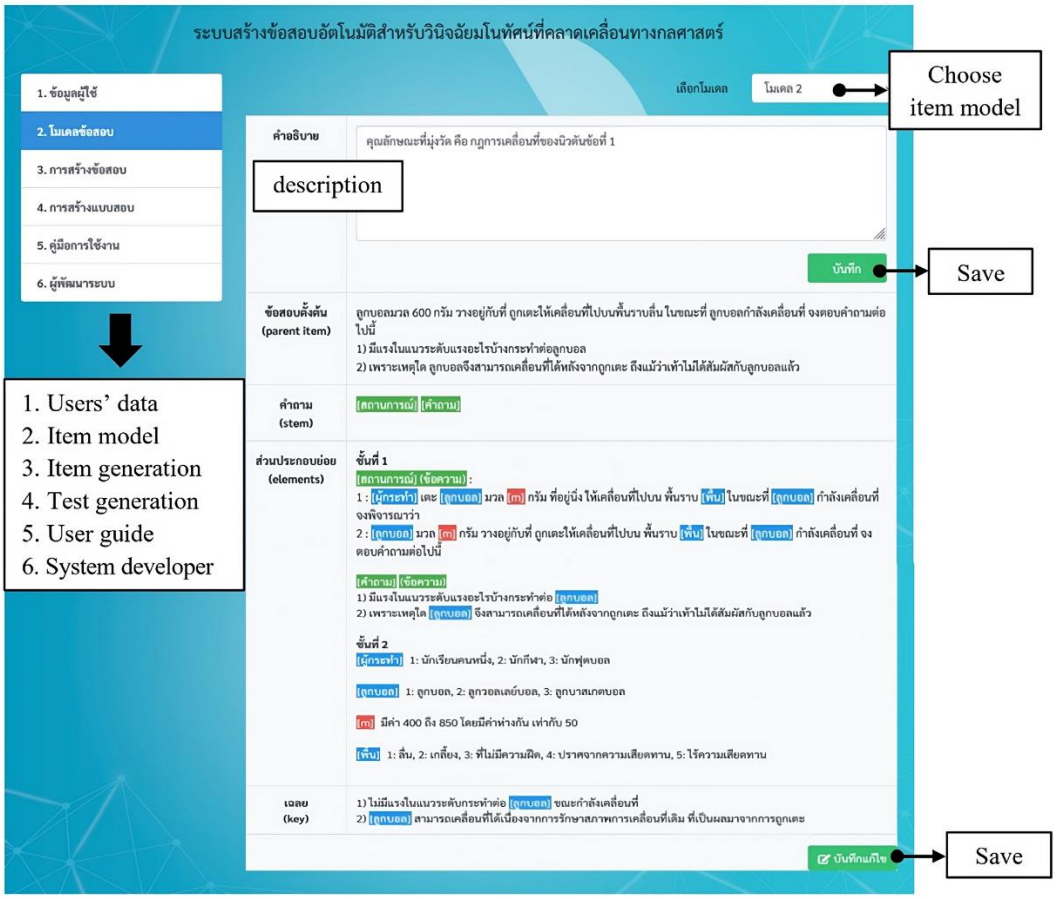

Figure 3. Examples of the item model menu bar (Source: Authors' own elaboration)

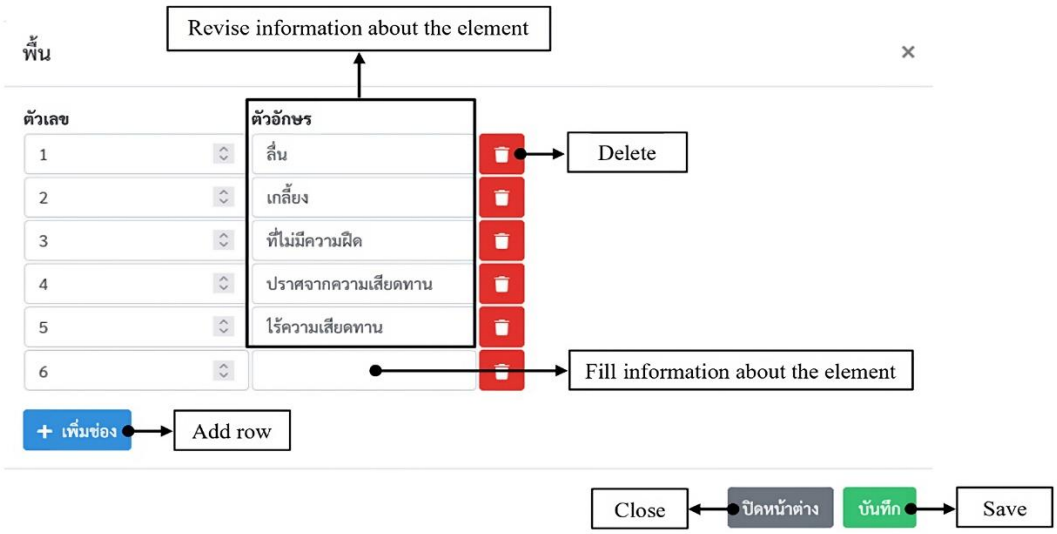

Figure 4. An example of the window after clicking on the element bar (Source: Authors' own elaboration)

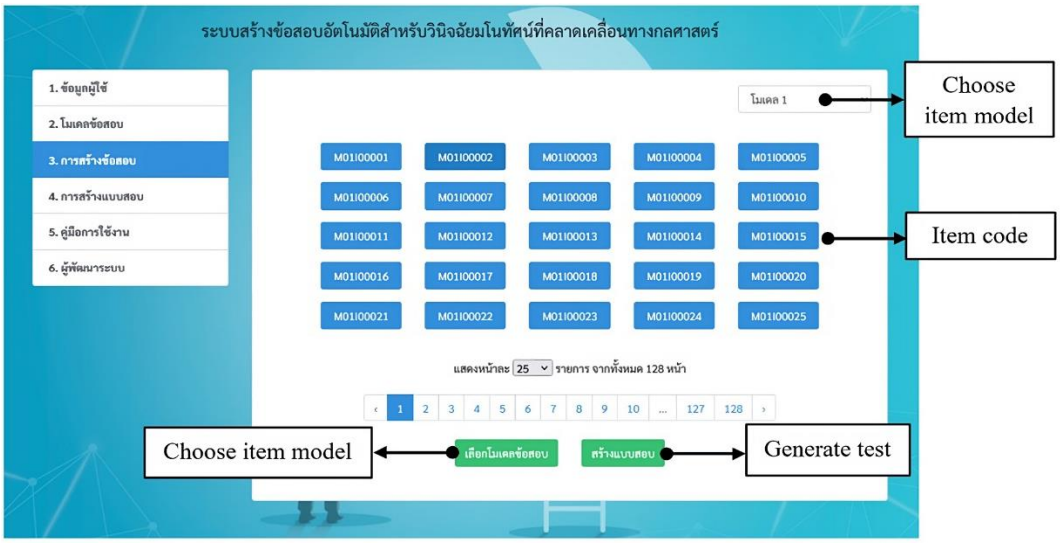

Figure 5. An example of the window after generating items (Source: Authors' own elaboration)

window will appear to allow users to revise the elements, as shown in **Figure 4**.

Item generation: The item generation menu bar shows all 18 item models, including the attributes of force and laws of motion intended to be assessed in each item model. Users can choose an item model they want to use to generate an item. After choosing an item model, the system will auto-generate items in each probable model by aggregating elements determined in the item model. Users can evaluate each test item generated in each item model by clicking the item code box, as shown in **Figure 5**.

Test generation: The test generation menu bar is used to organize the test form for the diagnosis of misconceptions about force and laws of motion that will be used. AIG system will randomly select items from each item model that has been chosen as the test form for the diagnosis. The details must be determined as follows:

1. The number of test forms to be generated.
2. The number of items randomly selected from each item model.
3. The two types of files of the test form, which are Word and PDF.

The test form for the diagnosis contains two parts, i.e., the test form with no keys and the test form with keys, which are on the other side of the test with no keys. The test form with keys is generated for scoring students' answers. While the test with no keys is generated for administering students.

For the generation of test forms with no keys, the details of the header can be added. Menu example for test generation is shown in **Figure 6**.

User guide: The user guide menu shows details about the usage of AIG system, which contains two important parts, i.e., the details of AIG system and the usage of AIG system. Users can download the user guide in the PDF form and can print it out.

Figure 6. Test generation menu bar (Source: Authors' own elaboration)

The quality of the user guide has been approved by experts in educational assessment and experts in information technology. It is found that the user guide for AIG system is high on quality when it comes to utility (M=4.46, SD=0.37) and feasibility (M=4.40, SD=0.23). Moreover, it is also highest on quality in terms of suitability (M=4.62, SD=0.36) and accuracy (M=4.57, SD=0.32). The user guide for AIG system is revised to ensure it responds to the needs of the users, who are high-school physics teachers.

System developer: The system developer menu shows details of the developer of AIG system for the diagnosis of misconceptions about force and laws of motion.

Usage for the admin: The usage of AIG system for the admin is similar to that for users in that there are two main parts, i.e., the login page and the main window for using AIG system. The admin can access and use the entire menu bars of AIG like users, but the users' data menu bar will be different for the admin. That is, the users' data menu bar is for managing users' information to identify the status of the users' usage. There are three status bars, i.e.,

(1) usage permission,

(2) usage suspension, and

(3) account deletion

Results of Verifying Quality of Automatic Item Generation System for Diagnosis of Force and Laws of Motion

Based on the results of the quality verifying of AIG system for the diagnosis of misconceptions about force and laws of motion by experts in educational assessment and in information technology, it is found that AIG system is highest on quality in all four dimension, which are

(1) utility (M=4.76, SD=0.37),

(2) feasibility (M=4.57, SD=0.24),

(3) propriety (M=4.71, SD=0.23), and

(4) accuracy (M=4.71, SD=0.37).

Based on the data on users' experiences in two dimensions, which are the pragmatic dimension and hedonic dimension, collected by physics instructors who tested AIG system. The data are used to improve AIG system in line with users' needs. The presentation is categorized according to the dimensions of users' experience, as follows.

Pragmatic dimension

The data on users' experiences with regards to the pragmatic dimension demonstrate the utility and capability of AIG system for the diagnosis of misconceptions about force and laws of motion. According to users, AIG system is useful and functional.

AIG system features a good design that makes it accessible and is user friendly. That is, AIG system is organized according to the operational sequence, which allows users to familiarize themselves with the system. They can look up the user guide to learn how to use the system by themselves. AIG system is not complicated; it only contains all the necessary menu bars that are well arranged and user-friendly. There are various menu bars that can be rearranged according to each user's preference. Each menu bar is also easily accessible. In addition, the icons are able to convey their meanings well, making the process of learning AIG system an easy one. Some menu bars are so easy to use that users don't need to look up the instructions in the user guide. More importantly, AIG system is able to process quickly. That is, with a click at a menu bar, AIG system will promptly show results. The design of AIG system is suitable for instructors that are not tech savvy.

AIG system is highly beneficial to physics instructors because it is responsive to the instructors' needs, especially when it comes to designing a large number of test forms that are different and will be used on students. It can reduce the chance of students copying answers, as they receive different tests. In addition, the system also reduces instructors' burden of generating test forms, as they can promptly come up with test forms on the diagnosis of misconceptions about force and laws of motion that cover content about force and laws of motion without having to revise the values in the item

model. The test forms on the diagnosis of misconceptions can be generated in both the Word format, which can be revised anytime, and in PDF format, which can be used on students right away.

Even though AIG system for the diagnosis of misconceptions about force and laws of motion is able to function as detailed above, it still has room for improvement to better respond to users' needs. Some of the improvements that can be made are detailed, as follows:

1. The rearrangement of the message box on the login page, which are
   (1) registration,
   (2) login, and
   (3) password retrieval.
   The colors of the login message box can be changed to differentiate it from other message boxes. This will ensure clarity and responsiveness.
2. The addition of the number next to the menu bars to inform users about the order of each menu.
3. The addition of the "save" button under the description header in the item model menu bar to save any revisions in the description. This action will separate the revisions in the description from all the revisions in the item model menu.
4. A diagnosis test with and without keys should be generated in one file to prevent a huge volume of test-form files and ensure usage leniency.

Hedonic dimension

The data on users' experiences in the hedonic dimension reflects interests and impressions of users to AIG system for the diagnosis of misconceptions about force and laws of motion. Interests and impressions can be summarized as follows:

AIG system is easily accessible, easy to use, fast-processing, and user-friendly. It also allows users to revise details according to their needs. In addition, AIG system also fits with users that need to generate test forms for the diagnosis of misconceptions about force and laws of motion in a short period of time.

DISCUSSION

AIG system for the diagnosis of misconceptions about force and laws of motion, which we developed has a development process that is in line with the three steps of the process of AIG. They are

(1) determining the content for the item model,

(2) generating the item model, and

(3) generating the items and assess the similarity of the items (Gierl & Lai, 2016; Graf et al., 2005).

That is, we started by determining the content to be used in generating items from the item model in the form of the cognitive model for AIG. The content for the item model was determined using the strong theory. The next step is generating an item model that is in line with the cognitive model for AIG. The item model we developed is a constructed-response item model. The item model is, thus, divided into two parts (i.e., stem and auxiliary information). Then, experts in physics teaching assess the content and the logic identified in the cognitive model for AIG and the structure of test items in the item model. There are three aspects of assessment (i.e., content, logic, and presentation). The suggestions would be used to improve the cognitive model for AIG and the item model. The item model can generate a large number of items that are of quality and reasonable. The final step is using AIG system for the diagnosis of misconceptions about force and laws of motion with all the aggregated elements identified in each item model. This results in a large number of meaningful items generated in a short period of time. The CSI is used to assess the similarity of the items generated from the same item model.

This research is aimed at generating item models to generate parallel items, with the intention of assessing the same attributes with details in the items that are not too similar. We, thus, developed an item model with an average CSI value ranging 0.60-0.80. The CSI is used to consider the repeated words in each item generated from the same item model. The item model with a high CSI average and a low SD will generate a large number of items that are similar with one another, also known as isomorphs or clones. Item models with a low CSI average and a high SD will generate items that are highly different from one another, also known as variants (Gierl & Lai, 2013). We used the study by Latifi et al. (2017), which found that the item model that generated isomorphs will have a CSI average of over 0.70 and a SD of lower 0.10. While item model that generated variants will have a CSI average of lower 0.70 and an SD of over 0.10. It offers an approach to develop an item model with CSI average ranging 0.70±0.10 to ensure the items generated are not too similar or are not too different. It should be noted that the determination of CSI average cutoff that will determine whether the item model will generate items that are isomorphs or variants is an interesting topic that could be picked up in the next research.

AIG system for the diagnosis of misconceptions about force and laws of motion that was developed is in line with the process of AIG system, which means it meets the four criteria of the standard evaluation determined by the Joint Committee on Standards for Educational Evaluation, which are

(1) utility,

(2) feasibility,

(3) propriety, and

(4) accuracy (Yarbrough et al., 2011).

This indicates AIG system is able to generate tests that are accurate and meaningful.

In addition, we also improved AIG system for the diagnosis of misconceptions about force and laws of motion using the two dimensions of users' experiences, which are

(1) pragmatic dimension and

(2) hedonic dimension.

The former is a pragmatic tool that responds to each individual's general behavioral target. It considers clarity, supportiveness, utility, and the ability to function and control. The latter is a sensory element that respond to each individual's behavioral target that leads to users' satisfaction. It considers distinction, impression, excitement and appeal (Hassenzahl, 2003; Krueger et al., 2020). The collection of users' data will help understand the needs and opinions of users who use AIG system. The data are important, as they can help improve AIG system to ensure it is highly responsive to users' needs. Moreover, we collected data on users' experiences by means of interviews to get in-depth data on the usage of AIG system and users' needs (Hussain et al., 2019).

CONCLUSIONS

The research results found that the cognitive model for AIG and every pair of item model are of quality. It was also found that AIG system for the diagnosis of misconceptions about force and laws of motion is of quality. Physics instructors and scholars can use AIG system to generate a huge volume of items aimed at diagnosing misconceptions about force and laws of motion in a short period of time. This is because AIG system, which we developed contains an item model for the diagnosis of misconceptions about force and laws of motion. Instructors and scholars must study the basic concepts of the model before using AIG system.

This development of AIG system for the diagnosis of misconceptions about force and laws of motion is an approach for developing AIG system that will respond to users' needs. The system must possess the following attributes, i.e., the system should be accessed through various browsers, the result display should be adjusted to suit different device screens, the system's menu bars should be organized according to the operational sequence to ensure it is easy to use. It should also contain only the necessary menu bars that are organized in such a way that make them coherent and understandable. The icons used should clearly reflects the actual function of the menus and allow users to revise the item model and all the necessary details according to their needs.

Author contributions: All authors have sufficiently contributed to the study and agreed with the results and conclusions.

Funding: This study was supported by the 100th Anniversary Chulalongkorn University for Doctoral Scholarship and the CU Graduate School Thesis Grant, Graduate School, Chulalongkorn University.

Ethical statement: Authors stated that the study did not require approval from an ethics committee. data was collected on user's experiences from small group of teachers who were not sensitive samples. Besides, questions were used to gather data were general questions in order to improve AIG system. Informed consents were obtained from the participants. Highest ethical guidelines were followed throughout the study.

Declaration of interest: No conflict of interest is declared by authors.

Data sharing statement: Data supporting the findings and conclusions are available upon request from the corresponding author.

REFERENCES

Aini, F. N., Sutopo, & Suyudi, A. (2021). Teaching integrated Newton's laws of motion for high school students. *AIP Conference Proceedings, 2330*, 050013. https://doi.org/10.1063/5.0043193

de la Torre, J., & Minchen, N. (2014). Cognitively diagnostic assessments and the cognitive diagnosis model framework. *Psicología Educativa* [*Educational Psychology*], *20*(2), 89-97. https://doi.org/10.1016/j.pse.2014.11.001

Embretson, S., & Yang, X. (2006). Automatic item generation and cognitive psychology. In C. R. Rao, & S. Sinharay (Eds.), *Handbook of statistics* (pp. 747-768). North Holland. https://doi.org/10.1016/S0169-7161(06)26023-1

Gierl, M. J., & Lai, H. (2013). Instructional topics in educational measurement (ITEMS) module: Using automated processes to generate test items. *Educational Measurement: Issues and Practice, 32*(3), 36-50. https://doi.org/10.1111/emip.12018

Gierl, M. J., & Lai, H. (2016). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice, 35*(4), 6-20. https://doi.org/10.1111/emip.12129

Gierl, M. J., & Lai, H. (2018). Using automatic item generation to create solutions and rationales for computerized formative testing. *Applied Psychological Measurement, 42*(1), 42-57. https://doi.org/10.1177/0146621617726788

Gierl, M. J., Zhou, J., & Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning and Assessment, 7*(2), 1-51.

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). *Psychometric and cognitive analysis as a basis for the design and revision of quantitative item models*. ETS. https://doi.org/10.1002/j.2333-8504.2005.tb02002.x

Gurel, D. K., Eryılmaz, A., & McDermott, L. C. (2015). A review and comparison of diagnostic instruments to identify students' misconceptions in science. *EURASIA Journal of Mathematics, Science and Technology Education, 11*(5), 989-1008. https://doi.org/10.12973/eurasia.2015.1369a

Hassenzahl, M. (2003). The thing and I: Understanding the relationship between user and product. In M. A. Blythe, K. Overbeeke, A. F. Monk, & P. C. Wright (Eds.), *Funology: From usability to enjoyment* (pp. 31-42). Kluwer Academic. https://doi.org/10.1007/1-4020-2967-5_4

Hussain, A., Hussein, I., Mkpojiogu, E. O., & Sarlan, A. (2019). The state of user experience design (UXD) practice in Malaysia: An in-situ interview approach. *International Journal of Innovative Technology and Exploring Engineering, 8*(8S), 498-505.

Javidanmehr, Z., & Sarab, M. R. A. (2017). Cognitive diagnostic assessment: Issues and considerations. *International Journal of Language Testing, 7*(2), 73-98.

Kaniawati, I., Fratiwi, N. J., Danawan, A., Suyana, I., Samsudin, A., & Suhendi, E. (2019). Analyzing students' misconceptions about Newton's laws through four-tier Newtonian test (FTNT). *Journal of Turkish Science Education, 16*(1), 110-122.

Krueger, A. E., Pollmann, K., Fronemann, N., & Foucault, B. (2020). Guided user research methods for experience design–A new approach to focus groups and cultural probes. *Multimodal Technologies and Interaction, 4*(3), 1-22. https://doi.org/10.3390/mti4030043

Lai, H., Gierl, M. J., Byrne, B. E., Spielman, A. I., & Waldschmidt, D. M. (2016). Three modeling applications to promote automatic item generation for examinations in dentistry. *Journal of Dental Education, 80*(3), 339-347. https://doi.org/10.1002/j.0022-0337.2016.80.3.tb06090.x

Latifi, S., Gierl, M., Wang, R., Lai, H., & Wang, A. (2017). Information-based methods for evaluating the semantics of automatically generated test items. *Artificial Intelligence Research, 6*(1), 69-79. https://doi.org/10.5430/air.v6n1p69

Narjaikaew, P. (2013). Alternative conceptions of primary school teachers of science about force and motion. *Procedia-Social and Behavioral Sciences, 88*, 250-257. https://doi.org/10.1016/j.sbspro.2013.08.503

National Research Council. (1997). *Science teaching reconsidered: A handbook*. National Academies Press.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Saglam-Arslan, A., & Devecioglu, Y. (2010). Student teachers' levels of understanding and model of understanding about Newton's laws of motion. *Asia-pacific Forum on Science Learning and Teaching, 11*(1), Article 7.

Sinharay, S., & Johnson, M. S. (2008). Use of item models in a large-scale admissions test: A case study. *International Journal of Testing, 8*(3), 209-236. https://doi.org/10.1080/15305050802262019

Sinharay, S., & Johnson, M. S. (2013). Statistical modeling of automatically generated items. In M. J. Gierl, & T. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp.183-195). Routledge.

Sornkhatha, P., & Srisawasdi, N. (2013). Supporting conceptual development in Newton's laws of motion using an interactive computer-simulated laboratory environment. *Procedia-Social and Behavioral Sciences, 93*, 2010-2014. https://doi.org/10.1016/j.sbspro.2013.10.157

Tang, R., Hu, Z., Henry, N., & Thomas, A. (2021). A usability evaluation of research data management librarian academy (RDMLA): Examining the impact of learner differences in pedagogical usability. *Journal of Web Librarianship, 15*(3), 154-193. https://doi.org/10.1080/19322909.2021.1937442

Thibaut, L., Ceuppens, S., De Loof, H., De Meester, J., Goovaerts, L., Struyf, A., Boeve-de Pauw, J., Dehaene, W., Deprez, J., De Cock, M., Hellinckx, L., Knipprath, H., Langie, G., Struyven, K., van de Velde, D., van Petegem, P., & Depaepe, F. (2018). Integrated STEM education: A systematic review of instructional practices in secondary education. *European Journal of STEM Education, 3*(1), 02. https://doi.org/10.20897/ejsteme/85525

Wancham, K., Tangdhanakanond, K., & Kanjanawasee, S. (2022). The construction and validation of the cognitive model of force and motion for a diagnosis of misconceptions. *Journal of Education Naresuan University, 24*(3), 60-70.

Wancham, K., Tangdhanakanond, K., & Kanjanawasee, S. (2023). Sex and grade issues in influencing misconceptions about force and laws of motion: An application of cognitively diagnostic assessment. *International Journal of Instruction, 16*(2), 437-456. https://doi.org/10.29333/iji.2023.16224a

Yarbrough, D. B., Shula, L. M., Hopson, R. K., & Caruthers, F. A. (2011). *The program evaluation standards: A guide for evaluators and evaluation users*. SAGE.

Yasri, P. (2014). A systematic classification of student misconceptions in biological evolution. *International Journal of Biology, 3*(2), 31-41. https://doi.org/10.20876/ijobed.06781