



# A Method for Generating Testlets

Mark J. Gierl<sup>1</sup><sup>a</sup> and Tahereh Firoozi<sup>2</sup><sup>b</sup>

<sup>1</sup>Measurement, Evaluation, and Data Science, Faculty of Education, University of Alberta, Edmonton, Alberta, Canada

<sup>2</sup>School of Dentistry, Faculty of Medicine and Dentistry, University of Alberta, Edmonton, Alberta, Canada

**Keywords:** Automatic Item Generation, Testlet Generation, Item Modelling.


**Abstract:** A testlet is a set of two or more items based on the same scenario. A testlet can be used to measure complex problem-solving skills that require a series or sequence of steps. A testlet is challenging to write because it requires one unique scenario and two or more items. Despite this challenge, large numbers of testlets are often required to support formative and summative computerized testing. The purpose of our study is to address the testlet item writing challenge by describing and demonstrating a systematic method that can be used to create large numbers of testlets. Our method is grounded in the three-step process associated with template-based automatic item generation. To begin, we describe a testlet-based item model. The model contains global and local variables. Global variables are unique to testlet generation because they can be used throughout the testlet, meaning that these variables can be used to place content anywhere in the testlet. Local variables, on the other hand, are specific to each item model in the testlet and can only be used in the same item model. Next, we present four cases that demonstrate how global and local variables can be combined to generate testlets. Each case provides a practical example of how the testlet item model can be used to structure global and local variables in order to generate diverse sets of test items. We conclude by highlighting the benefits of testlet-based automatic item generation for computerized testing.


## 1 INTRODUCTION

As the importance of technology in society continues to increase, countries require skilled workers who can produce new ideas, make new products, and provide new services. The ability to create these ideas, products, and services will be determined, in part, by the effectiveness of our educational programs. Students must think, reason, and solve complex problems in a world that is shaped by knowledge, information, and communication technologies (Auld & Morris, 2019; Chu et al., 2017). Educational testing has an important role to play in helping students acquire these skills. Educational tests, once developed to satisfy demands for accountability and outcomes-based *summative* testing, are now expected to provide teachers and students with timely, detailed, *formative* feedback to support teaching and learning. Formative principles can guide testing practices to help us meet these teaching and learning outcomes. Formative

principles can include any assessment-related activities that yield constant and specific feedback to modify teaching and improve learning, including administering tests and providing students with their scores more frequently (Black & Wiliam, 1998, 2010). But when testing and score reporting occur more frequently, more tests are required. These tests must be created efficiently, they must be economical, and they must adhere to a high standard of item development quality.

Fortunately, this requirement for frequent and timely educational testing coincides with the changes occurring in computer technology. Developers of educational tests are now implementing *computerized tests* at an extraordinary rate. For example, the world's most popular achievement test—the Programme for International Student Assessment (PISA) conducted by the Organization for Economic Cooperation and Development (OECD)—is now computerized. While the first five cycles of the test administered were paper-based, 58 of the 72 participating

<sup>a</sup> <https://orcid.org/0000-0002-2653-1761>

<sup>b</sup> <https://orcid.org/0000-0002-6947-0516>

countries in PISA 2015 took the computer-based version. Here is another milestone: 2016 marked the first academic year that US states administered more computer than paper-based tests (He & Lao, 2018). Provincial agencies in Canada, to cite another example, are also moving to computerized testing. Alberta Education introduced computerized tests in 2010 but limited use to the Grades 6 and 9 achievement testing programs. Computerized testing has now expanded to include all provincial tests including the achievement tests as well as the Grade 12 diploma examinations. These examples demonstrate that computerized testing has become the hallmark of 21st century assessment.

Computerized testing offers important benefits to support teaching and promote learning. Computers permit testing on-demand thereby ensuring students can write exams whenever they or their teachers believe feedback is required. Items on computerized tests are scored immediately thereby providing students with prompt feedback. Computerized tests are scored automatically thus reducing the time teachers would typically spend on grading. In short, computerized testing can help educators infuse formative principles into their testing practices.

Despite these important benefits, the advent of computerized testing has also raised noteworthy challenges, particularly in the area of test item development. Educators *must* have access to large numbers of diverse, high-quality test items to implement computerized testing because items are continuously administered. Hence, thousands of items are needed to develop the banks necessary for computerized testing. These banks must also be replenished to ensure that students receive a continuous new supply of content. Unfortunately, educational test items, as they are currently created, are time-consuming and expensive to develop because each individual item must be written by a content specialist and, when necessary, each individual item must be reviewed and revised. Hence, item development is one of the most important problems that must be solved before we can migrate to a broad computerized testing system that can guide formative and summative assessment because large numbers of high-quality, content-specific items are required (Andrade et al., 2019; Gierl et al., 2018; Leo et al., 2019; Karthikeyan et al., 2019; Morisson & Embretson, 2018).

## 2 OVERVIEW OF AUTOMATIC ITEM GENERATION

One method that can help address this item development challenge is *automatic item generation* (AIG) (Gierl & Haladyna, 2013; Gierl, Lai, & Tanygin, 2021; Irvine & Kyllonen, 2002). AIG is a rapidly evolving research area where cognitive and psychometric modelling practices are used to create items with the aid of computer technology. It can be used to address the challenging task of quickly and economically producing large numbers of high-quality, content-specific items.

Gierl and Lai (2016a, 2016b, 2017) described a three-step process for generating items. In step 1, a content specialist creates a cognitive model. A cognitive model is a representation that highlights the content required to generate new test items. In step 2, an item model is created. An item model is a template that specifies the variables in a test item that can be manipulated to produce new items. Item models provide the foundation for the AIG method described in our study. Hence, it is often referred to as template-based AIG (Gierl, Lai, & Tanygin, 2021). In step 3, algorithms are used to place the cognitive content into the item model. Taken together, this process can be used to generate hundreds of items from a single item model.

### 2.1 Multiple-Choice Item Model

The selected-response item format, generally, and the single-answer, multiple-choice item, specifically, is the most common format used in educational testing (Clauser et al., 2006; Daniel et al., 2019; Rencic et al., 2016). Students write thousands of multiple-choice items over the course of their academic lives. A multiple-choice item contains a stem and options and, at times, auxiliary information. The stem contains context, content and/or the question the student is required to answer. The options include a set of alternative answers with one correct and two or more incorrect options or distractors. Auxiliary information includes any additional context, either in the stem or options, required to answer the item, including images, tables, graphs, diagrams, audio, or video. The popularity of this item format can be attributed to its many well-documented strengths. The single-answer, multiple-choice item is easy to score. The ease of scoring means that it yields highly reliable test score results. It can be used to measure a broad range of content. It can be used to measure a range of different knowledge and skills. It can be administered and scored electronically. The single-answer, multiple-

choice item also has two important weaknesses. The student must spend a substantial amount of time and effort to read and integrate the content in the stem to answer a single item. In addition, a single-answer, multiple-choice item cannot measure either the series or the sequence of steps required to solve a complex problem because it is limited to a single item.

To address these limitations, alternative item formats can be used. One format that has the strengths of a single-answer, multiple-choice item and that also overcomes its weaknesses is called a testlet. A testlet—also referred to as a context-dependent item set, item bundle, or case-based item set—is a set of two or more items based on the same scenario, prompt, or vignette (Wainer & Kiely, 1987). Because a testlet contains a scenario with two or more items, it can be used to measure more complex problem-solving skills that require, for example, a series or sequence of steps (Bradlow et al., 1999). The testlet is administered as a unit where students must answer the items in the set using the same scenario. This item format has many benefits for assessing complex reasoning and problem-solving skills. For example, the testlet is effective for assessing a student's ability to synthesize and sort multiple sources of relevant and irrelevant data. It is easy to score. It yields highly reliable test score results. It can be used to measure a broad range of content and context-specific information. It can be administered and scored electronically. It reduces testing time because one scenario is associated with two or more items rather than a single item (Bradlow et al., 1999; Min & He, 2014; Wainer & Kiely, 1987). But the most important benefit of a testlet is that it can be used to measure a broad range of knowledge and skills as well as complex knowledge and skills compared to the single-answer, multiple-choice item because the content in the scenario is linked to more than one item. As a result, the item set can be used to measure both the outcomes and the sequence or steps required for complex problem-solving tasks. The scenarios are also content specific. Hence, many diverse scenarios can be assessed using a different item set with different contexts. In short, a testlet has all of the benefits of single-answer, multiple-choice items with the added benefit of allowing examiners to assess complex critical reasoning skills within a context-dependent scenario.

Despite these important benefits, the testlet has one important disadvantage. It is very time consuming to create. Each testlet requires one unique scenario and two or more items. The content specialists must be familiar with both the context and the reasoning process in order to create the scenario

in the stem and the associated items in the set. While a considerable literature exists on writing single-answer, multiple-choice items, there is virtually no literature on how to write a testlet (Lane et al., 2016). Hence, the complexity of the item writing task is compounded by the lack of guidance on how to execute the task. Finally, large numbers of testlets are required to support formative and summative assessments. A formative assessment designed to measure complex reasoning must focus on each step in the task in order to provide students with feedback in the form of a scored outcome. To ensure that the skills are generalizable, the formative assessment must also measure reasoning across a broad range of scenarios. A testlet can help students develop strong reasoning skills by providing opportunities for ample practice and feedback. This practice and feedback should be conducted using many different contexts. These requirements demonstrate that to implement a formative assessment focused on reasoning skills, large numbers of context-specific testlets are needed, where students received their score on each item.

Large numbers of testlets are also required for summative assessments. The purpose of administering a summative assessment is to measure complex reasoning and to ensure that students can demonstrate they are competent in implementing this important skill. Because the testlet provides an effective method for measuring complex reasoning, some high-stakes summative assessments only use this item format. For example, the Australian Dental Council—which is the accreditation authority for all Australian dental professions—administers its written summative exam to dentists as a computer-delivered multiple-choice test that only contains testlets. Examinees must complete 56 clinical scenarios, each with five related items, for a total of 280 items. Each item has one correct answer as it relates to the information in the clinical scenario. This example demonstrates that implementing a complex reasoning summative assessment requires large numbers of context-specific testlets.

In short, the most important challenge that must be addressed when creating and implementing testlet-based assessments resides with the item development process. Examiners must have access to large numbers of diverse, context-specific testlets that contain high-quality test items. Hundreds of testlets with, potentially, thousands of items are needed to implement formative and summative assessments. Unfortunately, testlets are time consuming and expensive to develop because each individual scenario and its associated items must be written and reviewed by a content specialist. The task of writing

thousands of items that can measure the knowledge and skills presented in hundreds of scenarios—without well-established item development guidelines—is a formidable task.

## 2.2 Testlet Item Model

A testlet can be created as part of the three-step AIG process described by Gierl and Lai. A testlet can be formatted as a type of template. Therefore, testlets can be generated by expanding the item model in Step 2. We begin by describing important concepts required to generate testlets within a template-based AIG structure. Then, we illustrate how these concepts can be used to generate a scenario with three associated items thereby creating a bank that contains many three-item testlets. We use a very simple and intuitive example to demonstrate our testlet generation method. A testlet is composed of one overarching scenario and two or more items that are linked to the content in the scenario. Separate item models are positioned after the scenario and these item models are used to generate each item in the testlet. We use three different item models in our example to generate three unique items per testlet. Hence, we need to create one scenario and three item models for our testlet example.

The foundational concept that guides testlet-based AIG is the implementation and differentiation of global and local variables. Variables in AIG contain values. Values contain the content that will be varied to create new test items. Global variables are typically introduced in the scenario. Global variables can be used throughout the testlet, meaning that these variables can be used anywhere in the testlet to vary the content (i.e., in the scenario as well as in one or more of the item models). Global variables are very flexible. They can be related to other global variables as well as to all of the local variables. The relations between global to global variables are maintained throughout the testlet. The relations between the global to local variables are only maintained for one specific item model. Global variables, therefore, permit the content specialist to link the scenario to an item model as well as to link one item model to another item model. Global variables are denoted with the letter G. The values associated with the global variable are denoted by numbers in brackets that range from 1 to the total number of values, n.

Local variables are specific to one item model. Local variables can be related to other local variables in the same item model. However, local variables cannot be related to global variables or to local variables in other item models. In other words, a local

variable functions in the same way as a variable functions when creating a single-answer, multiple-choice item model (see Gierl, Lai, & Tanygin, 2021, Chapter 2). Or said differently, a single-answer, multiple-choice item model only contains local variables. Local variables permit the content specialists to add unique content into each item model so that each model is unique and independent from the other models in the testlet. Local variables are denoted with the letter L. The values associated with the local variable are denoted by numbers in brackets that range from 1 to the total number of values, n.

Next, we present four different testlet-based AIG cases. Each case serves as a demonstration of how global and local variables can be used to generate testlets, where each testlet contains a unique three-item set.

### 2.2.1 Case 1: Scenario, Global Variable

Case 1 is shown in Figure 1. It contains the scenario. The scenario includes the context used to structure the testlet. All items in the testlet are based on this scenario. The scenario can only include global variables. Our example includes one global variable. The global variable is denoted as G1 in this example. The global variable CITY (G1) contains three values (1-3): London [G1(1)], Paris [G1(2)], and Melbourne [G1(3)]. CITY is a global variable that appears in three of the four cases.

Gregory is traveling to **G1:CITY**. He is a 30-year-old male from Canada. Gregory is travelling alone.

#### **CITY [GLOBAL]:**

G1(1) London

G1(2) Paris

G1(3) Melbourne

Figure 1: Case 1: Scenario with 1 global variable.

### 2.2.2 Case 2: Item Model 1, Local Variable

Case 2 is shown in Figure 2. It contains the first item model and one local variable. The local variable, TIME (L1), is specific to this model, meaning that TIME is used in item model 1 but in no other item models in this testlet. TIME contains three values: in the morning [L1(1)], early in the afternoon [L1(2)], and in the evening [L1(3)]. Each item model also contains a list of the correct and the incorrect options. In our example, item model 1 includes nine correct options. It also contains nine incorrect options an

He arrived **L1:TIME** and is really hungry.  
What should he have to eat?

**TIME [LOCAL]:**  
L1(1) in the morning  
L1(2) early in the afternoon  
L1(3) in the evening

**CORRECT OPTIONS:**  
Full english breakfast [G1(1), L1(1)]  
High tea [G1(1), L1(2)]  
Bangers and mash [G1(1), L1(3)]  
Croque madame [G2(1), L1(1)]  
Crepes [G2(2), L1(2)]  
Cassoulet [G2(3), L1(3)]  
Vegemite on toast [G3(1), L1(1)]  
Lamingtons [G3(2), L1(2)]  
Barbecued snags [G3(3), L1(3)]

**INCORRECT OPTIONS:**  
Full english breakfast  
High tea  
Bangers and mash  
Croque madame  
Crepes  
Cassoulet  
Vegemite on toast  
Lamingtons  
Barbecued snags

Figure 2: Case 2: Item model 1 with 1 local variable.

because the same list is used, meaning that when option in the list is not correct, then that option can be used as a distractor. This strategy for selecting incorrect options is common in template-based AIG (Gierl et al., 2012). Notice that each correct option is constrained by the local variable in the item model. For example, when the CITY is London [G1(1)] and the TIME is in the morning [L1(1)] the correct option is Full english breakfast [G1(1), L1(1)].

### 2.2.3 Case 3: Item Model 2, Global Variable

Case 3 is shown in Figure 3. It includes the second item model in our example. Like Case 2, Case 3 only contains a single variable. However, unlike Case 2, the variable in Case 3 is our global variable, CITY. Recall that global variables can be used anywhere in the testlet. Item model 2 includes three correct options. The three correct options along with five additional incorrect options are used to create the incorrect option list. Each correct option is constrained by the global variable in the item model.

For example, when the CITY is London [G1(1)] the correct option is Pound [G1(1)].

### 2.2.4 Case 4: Item Model 3, Global and Local Variable

Case 4 is presented in Figure 4. It contains the third item model. Case 4 is the most complex model in our example because it contains both a global and a local variable. In total, our global variable CITY is used three times (Cases 1, 3, and 4). This model also contains a local variable NEW LOCATION (L3) that is unique to item model 3. The NEW LOCATION local variable is not used in any other model in our example. Case 4 includes one global variable with three values [G1(1-3)] and one local variable with three values [L3(1-3)]. It also includes two correct options. Each correct option is constrained by both the global and local variables. For example, train as a correct option can be used to travel from London to Berlin [G1(1), L3(1)], from Paris to Berlin [G1(2), L3(1)] or from Sydney to Melbourne [G1(3), L3(3)]. The incorrect options list contains four—admittedly weak—implausible answers in our example.

After a few days, Gregory needs more currency. He is at the currency exchange booth. Which currency should he get for use in **G1: CITY**?

**CITY [GLOBAL]:**  
London G1(1)  
Paris G1(2)  
Melbourne G1(3)

**CORRECT OPTIONS:**  
Pound [G1(1)]  
Euro [G1(2)]  
Dollar [G1(3)]

**INCORRECT OPTIONS:**  
Pound  
Euro  
Dollar  
Pesos  
Sols  
Ruble  
Franc  
Forint

Figure 3: Case 3: Item model 2 with 1 global variable.



After a couple of weeks, Gregory gets tired of **G1: CITY** and decides to travel to **L2: NEW LOCATION**. How should he get there?

**CITY:**  
 London G1(1)  
 Paris G1(2)  
 Melbourne G1(3)

**NEW LOCATION:**  
 Berlin (can start from London or Paris) L3(1)  
 Sydney (can only start from Melbourne) L3(2)  
 Los Angeles (can start from any city) L3(3)

**CORRECT OPTIONS:**  
 By train [G1(1), L3(1); G1(2), L3(1); G1(3), L3(2)]  
 By air [G1(1), L3(3); G1(2), L3(3); G1(3), L3(3)]

**INCORRECT OPTIONS:**  
 By foot  
 By horse  
 By scooter  
 By boat

Figure 4: Case 4: Item model 3 with 1 global and 1 local variable.

### 2.3 Importance of Global Variables in Testlet-Based AIG

Four cases were presented to demonstrate the important role that global variables play in testlet-based AIG. Local variables are familiar to AIG researchers and practitioners. Local variables are used to generate single-answer, multiple-choice items. In this context, they are simply called variables. The creation and implementation of local variables are well documented in the AIG literature (Gierl, Lai, & Tangin, 2021, Chapter 3). Global variables, on the other hand, serve as a new modelling concept that is unique to testlet-based AIG. Global variables are typically introduced with the scenario. This strategic decision means that these variables can be used in any item model thereby allowing the content specialists to link the scenario to the item models. CITY was a global variable used frequently in our examples. This global variable appeared in Cases 1, 3, and 4. Global variables are very flexible. They can be used in the scenario (Case 1), they can be used as stand alone variables in an item model (Case 3), and they can be used with local variables in an item model (Case 4). The

main function of a global variable is that it can link the scenario to the item models and it can also link the item models with one another. In short, global variables play a very important role in testlet-based AIG because they provide the content specialist with both the flexibility and freedom to create dependencies throughout the testlet. An example of one testlet generated using the content from the models in Figures 1 to 4 is shown in Figure 5.

Gregory is traveling to Paris. He is a 30-year-old male from Canada. Gregory is travelling alone.

He arrived in the evening and was really hungry. What should he have to eat?

- Full english breakfast
- Cassoulet
- Croque madame
- Lamingtons

After a few days, Gregory needs more currency. He is at the currency exchange booth. Which currency should he get for use in Paris?

- Euro
- Dollar
- Pound
- Pesos

After a couple weeks, Gregory gets tired of Paris and decides to travel to Berlin. How should he get there?

- By foot
- By car
- By horse
- By scooter

Figure 5: A sample testlet containing one scenario and three items.

## 3 CONCLUSIONS

AIG is an item development method where cognitive and psychometric models are used to produce test items with the support of computer technology. AIG can be used to create large numbers of items. Large numbers of items are often required to measure the thinking and reasoning skills that students need to implement when solving complex, real-world problems. Items that measure complex problem-

solving skills can be used to guide our practice and feedback strategies when implementing formative assessments and to evaluate competency and proficiency when implementing summative assessments. Traditional single-answer, multiple-choice items—while common in educational testing—cannot be used to measure the steps required to solve complex, often context-dependent, problems because this item format is limited to a single item.

Testlets can be used to overcome this limitation. A testlet is a set of two or more items based on the same scenario. Testlets are effective at measuring complex problem-solving skills because they include a set of items related to a common scenario thereby ensuring that different aspects or components of the problem—as it relates to the scenario—can be evaluated. Unfortunately, testlets are challenging to write. And large numbers of testlets are required for both formative and summative assessments.

To address this item development problem, testlets can be integrated into the three-step AIG process by creating a testlet item model. A testlet item model is unique because it contains two types of variables. Global variables can be used throughout the testlet to vary the content of the generated items. Global variables are typically introduced in the scenario so they can be used to link the content in the item model to the content in the scenario. Global variables can also be used to link content across two or more item models. Local variables are also used in testlet-based AIG. A local variable is specific to each item model and therefore cannot be used throughout the testlet. A local variable is used to help ensure the content in each item model in the testlet is unique. We provided four illustrative cases to demonstrate how global and local variables can be used independently or combined with one another to generate items.

To conclude, testlet-based AIG is a new method for scaling the item development process. It allows content specialists to create sets of items linked to a common scenario. The context for the scenario is limitless meaning the scenario can be short or long, it can contain a small or a large amount of content, it can contain a small or a large number of global variables, and it can be in any content area. In other words, the scenario in a testlet can be created to accommodate any problem-solving situation. Testlet-based AIG can also be used to measure a range of knowledge and skills because the length of the item set is flexible. A testlet can contain a small (e.g., 2) or a large number of items (e.g., >5) thereby allowing the content specialist to measure many different types of knowledge and skills as they relate to the content in the scenario. Finally, testlet-based AIG is

embedded within a well-established item development framework associated with AIG (Gierl & Lai, 2016b). This framework is structured using a three-step process, where the testlet item model is created in step 2. The framework also includes a method for validating the content thereby ensuring the generated items in the testlet accurately measure the intended curricular and cognitive outcomes on the computerized test of interest.

## REFERENCES

- Andrade, H., Bennett, R., & Cizek, G. (2019). *Handbook of Formative Assessment in the Disciplines*. Boca Raton, FL: CRC Press.
- Auld, E. & Morris, P. (2019). The OECD and IELTS: Redefining early childhood education for the 21st century. *Policy Futures in Education, 17*, 11-26.
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice, 5*, 7-74.
- Black, P. & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*, 81-90.
- Bradlow, E. T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168.
- Chu, S., Reynolds, R., Notari, M. & Lee, C. (2017). *21st Century Skills Development through Inquiry-Based Learning*. New York: Springer.
- Clauser, B. E., Margolis, M. J., & Case, S. M. (2006). Testing for licensure and certification in the professions. *Educational Measurement, 4*, 701-731.
- Daniel, M., Rencic, J., Durning, S. J., Holmboe, E., Santen, S. A., Lang, V., ... & Gruppen, L. D. (2019). Clinical reasoning assessment methods: a scoping review and practical guidance. *Academic Medicine, 94*(6), 902-912.
- Gierl, M. J., Bulut, O., & Zhang, X. (2018). Using computerized formative testing to support personalized learning in higher education: An application of two assessment technologies. In R. Zheng (Ed.), *Digital Technologies and Instructional Design for Personalized Learning* (pp. 99-119). Hershey, PA: IGI Global.
- Gierl, M. J., & Haladyna, T. (2013). *Automatic Item Generation: Theory and Practice*. New York: Routledge.
- Gierl, M. J. & Lai, H. (2016a). Automatic item generation. In S. Lane, M. Raymond, & T. Haladyna (Eds.), *Handbook of Test Development* (2<sup>nd</sup> edition, pp. 410-429). New York: Routledge.
- Gierl, M. J. and Lai, H. (2016b). A process for reviewing and evaluating generated test items. *Educational Measurement: Issues and Practice, 35*, 6–20.
- Gierl, M. J. & Lai, H. (2017). The role of cognitive models in automatic item generation. In A. Rupp & J. Leighton

- (Eds.), *The Wiley Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (pp. 124-145). New York: Wiley.
- Gierl, M. J., Lai, H., & Tanygin, V. (2021). *Advanced Methods in Automatic Item Generation*. New York: Routledge.
- Gierl, M. J., Lai, H., & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education*, *46*, 757-765.
- He, D., & Lao, H. (2018). Paper-and-pencil assessment. *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation* (pp. 1199-1200). Thousand Oaks, CA: Sage.
- Irvine, S. H., & Kyllonen, P. C. (2002). *Item Generation for Test Development*. Hillsdale, NJ: Erlbaum.
- Karthikeyan, S., O'Connor, E., & Hu, W. (2019). Barriers and facilitators to writing quality items for medical school assessments. *BMC Medical Education*, *19*:123.
- Lane, S., Raymond, M., & Haladyna, T. (2016.). *Handbook of Test Development* (2<sup>nd</sup> edition). New York: Routledge.
- Leo, J., Kurdi, G., Matentzoglou, N., Parsia, B., Sattler, U., Forge, S., Donato, G., & Dowling, W. (2019). Ontology-based generation of medical, multi-term MCQs. *International Journal of Artificial Intelligence in Education*, *29*, 145-188.
- Min, S., & He, L. (2014). Applying unidimensional and multidimensional item response theory models in testlet-based reading assessment. *Language Testing*, *31*(4), 453-477.
- Morrison, K. & Embretson, S. (2018). Item generation. *The Wiley Handbook of Psychometric Testing: A Multidisciplinary Reference on Survey, Scale and Test Development* (pp. 75-94). New York: Wiley.
- Rencic, Joseph & Durning, Steven & Holmboe, Eric & Gruppen, Larry. (2016). *Understanding the Assessment of Clinical Reasoning*. 10.1007/978-3-319-30064-1\_11.
- Wainer, H., & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*(3), 185-201.