International Journal of Assessment Tools in Education
2022, Vol. 9, Special Issue, 6–18
https://doi.org/10.21449/ijate.1196701
Published at https://ijate.net/ https://dergipark.org.tr/en/pub/ijate *Original article*

Creating multiple-choice items for testing student learning

Thomas Haladyna[1,*]

[1]Arizona State University, Arizona, USA

ARTICLE HISTORY
Received: Sep. 10, 2022
Accepted: Sep. 28, 2022

Keywords:
Multiple choice items,
Classroom assessment,
Measuring student learning.

Abstract: The use of multiple-choice items for classroom testing is firmly established for many good reasons. The content any unit or course of study can be well sampled. Test scores can be reliable (trusted). And time spent administering and scoring can be minimized. This article provides a current review of best practices in the design and use of a variety of multiple-choice formats for classroom assessment of student learning. One of the most serious problems facing current educators is developing test items that measure more than simple factual recall. It is important to measure understanding, comprehension, critical thinking, and problem solving. Not only are these types of higher-level thinking described, but items are presented that illustrate how this is done. These best practices are continually evolving. The objective is always to use tests to measure validly what students have learned as well as help students learn what they have not yet learned. We call this formative and summative assessment. Guidelines are presented showing good and bad practices. What may be surprising to readers is the extensive variety of formats and methods for gathering or generating new test items. Readers are encouraged to experiment with these formats. Some formats can very efficiently measure what students were supposed to learn.

1. INTRODUCTION

This article concerns the writing and use of multiple-choice (MC) test items for evaluating student learning in a classroom or course of study. Best practices are described that lead to the development and validation of MC items that can be part of an inventory of test items for formative and summative evaluation of student learning. Formative refers to the use of test items to help students learn. Think of formative as practice and feedback on how much learning has occurred. Summative refers to a piece of evidence used with other evidence to assign a student grade.

This article is based my extensive study of the origin of item development, related research, and considerable experience both with testing programs and testing in the classroom in elementary and secondary schools and in universities and professional schools. Much of the background for this article comes from references provided at the end of this article.

*CORRESPONDING AUTHOR: Thomas Haladyna ✉ tmh@asu.edu Arizona State University, Arizona, USA

e-ISSN: 2148-7456 /© **IJATE** 2022

1.1. Objectively-scorable Items

To start, the term *MC item* is limiting. A better term is objectively-scorable item (OSI). In this section, a family of OSIs is introduced and illustrated, one of which is the MC item format. Each is introduced, illustrated, and a brief comment is offered. The most comprehensive source of information about OSIs is in our book (Haladyna & Rodriguez, 2013).

1.1.1. *Conventional MC (CMC)*

The most basic OSI has a stem and three choices. Note that traditionally, the CMS has consisted had four or five choices. Extensive research over many years has led us to conclude that three choices are sufficient (Haladyna, Raymond, & Stevens, 2019; Rodriguez, 2016). These fourth and fifth choices typically fail to discriminate or are so implausible that no student would choose it. Creating fourth and fifth options is usually a waste of time for the item writer.

Which type of travel from Ankara to Istanbul is most economical? (STEM)

A. Bus (Correct)*

B. Train (Distractor)

C. Plane (Distractor)

The biggest objection to this recommendation is that guessing might influence the accuracy of a test score. A student might be a lucky or unlucky guesser. A recent review and study show that guessing is overrated as threat to the accuracy of any test score (Haladyna, submitted for publication). The average score for random guessing on a three-option item test is 33%. One would have to be extremely lucky to get a score much higher than deserved.

1.1.2. *Alternate choice (AC)*

Which type of travel is slowest on a trip from Ankara to Istanbul?

A. Personal auto

B. Bus

This format is most useful for higher-achieving students who ordinarily can narrow down any OSI test item to one or two plausible choices. Also, this item does not take up space and much reading time. The more items in a test, the more reliable the test score will be. With more items, you can cover more content as well.

An AC item can be modified under some circumstances in this way:

Which type of travel from Ankara to Istanbul is considered very economical?

A. Personal auto

B. Bus

C. Both A and B

D. Neither A nor B

This modification transforms the item into a four-option CMC.

1.1.3. *True-false (TF)*

This format has a questionable reputation that has limited it usefulness. However, stating a set of 30 declarative statements is convenient, about half of which are true and half of which are false. The efficiency of TF format is unmatched. Items are easy to write. Administration time is short. Scoring is easy. Reliability of test scores can be very high. If the items represent student learning outcomes, the test score is accurate.

The main problem with the TF format is the tendency to test factual recall instead of higher types of learning. Another criticism is random guessing. We must recognize that the floor of

the test score scale is 50%, so performance at this level would show a lack of student learning. Our standards for evaluating student performance should be much higher than 50%.

Mark A if true and B if false.

1. *Ankara is west of Istanbul.*
2. *Istanbul has a greater population compare with Ankara.*
3. *The climate of Ankara is generally warmer than Istanbul.*
4. *The climate of Ankara is rainier than Istanbul.*
5. *Rahmi M. Koç Museum is one of the best tourist attractions in Istanbul.*
6. *Gülhane Park is on the grounds of the Topkapi Palace.*

1.1.4. *Multiple true-false (MTF)*

The MTF format is useful for testing a family of related characteristics or examples of a concept. The MTF format has a lead question or an open-ended statement followed by a list of choices. Each choice is marked true or false by the student.

Which of the following are true regarding travel from Ankara to Istanbul?

1. *Bus travel is very slow.*
2. *Air travel is the most expensive.*
3. *Bus travel is the least expensive of all options.*
4. *Train travel is the most comfortable.*
5. *Train travel to Istanbul leaves you in the city center.*
6. *Car travel through Bursa is slower and more scenic.*

Like TF items, the MTF is very easy to write, administer, and score. Test scores can be very reliable. As with TF, we have the annoying interference of lucky or unlucky guessing, but as previously noted, random guessing is overrated as a threat. A major limitation of the MTF is a tendency to focus content narrowly instead of broadly. The examples in the above item deal with travel between two cities.

1.1.5. *Matching*

This format is underutilized. Little research exists on its use. Nonetheless, it should not be left out of your collection of OSI formats. As you can see, matching has one set of choices and many stems. So, it is an efficient type of MC.

A. Ankara	*1. The most populous city*
B. Istanbul	*2. The national capital*
C. Izmir	*3. The fastest growing city*
D. Bursa	*4. The most beautiful city of the four listed above*
	5. The most popular city for visitors
	6. Near the sea of Marmara

Many more items can be added using the same four choices above. This format is very efficient regarding administration. Also, items are easier to write. Finally, test scores tend to have high reliability because many items are used.

1.1.6. *Extended matching*

The extended matching format is used in situations where many choices are available. The example shown below comes from a medical test of cardiovascular symptoms and signed.

Options:

A. Radiofemoral delay

B. Pan-systolic murmur

C. Systolic blood pressure of 220 mmHg

D. Tapping apex beat

E. Chest pain eased by glyceryl trinitrate in five minutes

F. Third heart sound

G. Splinter hemorrhages

H. Breathlessness eased by lying flat

I. Slow-rising carotid pulses

J. Bardycardia with pulse rate 20 per minute

K. Chest pain eased by glyceryl trinitrate after an hour

Which of the choices above best describes the patients below?

1. *65-year-old man collapsed when running. He has a sustained heaving apex beat that is slightly displaced and an ejection systolic murmur.*
2. *An 80-year- old woman has an excruciating pain between the shoulder-blades. You palpate the right radial pulse but not the left.*
3. *A 70-year-old man had a myocardial infarction two years ago. He now has gradually increasing breathlessness worse on lying flat with crepitations in the lung bases.*
4. *A 65-year-old woman has been increasingly breathless over the last few years. On an auscultation, she has a loud first heart sound and a mid-diastolic murmur.*
5. *A 60-year-old man who smokes 20 cigarettes per day. He complains of a tight pain in the center of his chest, which comes on when he walks up stairs.*

With more items, all the choices listed above can have associated stems. The benefit of the extended matching is the wide coverage given to a variety of problems involving the heart.

1.1.7. *Testlets*

The testlet is the most useful and desirable of all OSI formats. It is often used in tests that measure reading comprehension. A short vignette or story is introduced and a series of items follows that provide indications of the student's comprehension. In science, an experiment or scientific observation is presented. Then a series of items follows that comprise the testlet. In teaching statistics to graduate students, my tests were designed around problems where a statistical procedure was applied. A set of generic CMC test items was used. All I had to do was change values of the problem to generate a new testlet. This approach to writing items and testlet has grown more popular now that automated item generation is a reality (See Gierl & Haladyna, 2015). Testlets are widely used in virtually all testing situations where complex use of knowledge and skills is required.

An excellent source of examples of testlets can be found on the following website:

https://www.act.org/content/act/en/products-and-services/the-act/test-preparation/reading-practice-test-questions.html?page=0&chapter=0. A Google search of testlets will yield a wealth of examples.

Testlets are typically more than a page in length, so one will not be presented here. However, understanding the structure is important, so a skeletal version of a testlet is presented in Table 1.

Table 1. *A sample skeleton version of a testlet*

Planning a Family Vacation
Passage: You are planning travel for your family from Ankara to Istanbul for a four-day vacation. Your father and mother trust you to provide useful information in planning this exciting trip. For each item, pick the correct response.
Items (only stems are provided)
1. How many miles is the distance from Ankara to Istanbul?
2. Which type of transportation is least costly to travel
3. What kind of climate might expect for this time of year in Istanbul?
4. How long will the trip be if we travel by car?
5. How expensive is air travel?
6. What is the cost of train travel?

With any testlet, any OSI format can be used. Also, the number of items can be quite long for each testlet. I once observed an entire test consisting of one testlet involving a group of teenagers going to a fair in their village.

1.1.8. *Completion Items*

There is one OSI that has no choices. It is the simplest of the family of OSI formats. The completion item is simply a question or prompt where a correct answer or performance is noted.

What is the most valuable natural resource of Turkey?

About how many hours is a train trip from Ankara to Istanbul?

With the completion item, a single right answer or a small set of right answers exists. The completion item is often used for measuring skills. This is an application of an item format for performance.

1.1.9. *Complex MC*

Here is one format is that not recommended. It looks like MC but that combinations that make it more challenging. The strike-out shows that this item type should NEVER be used.

~~*Which of the following modes of transportation are very slow?*~~

~~*1. Bus*~~

~~*2. Car*~~

~~*3. Walking*~~

~~*A. 1 and 2*~~

~~*B. 1 and 3*~~

~~*C. 2 and 3*~~

~~*D. 1, 2, and 3*~~

There are many bad variations of this format (sometimes called Type K). This format is widely rejected. With the exception of the complex MC, the other OSI formats have attractive features that recommend their use.

1.2. Validity and Reliability

The most important concept in the measurement of student learning is validity. We have extensive discussions of validity in various sources. For the sake of brevity, these principles are address validity. This brief section is intended to provide more context for choosing and using OSI formats for measuring student learning.

Validity refers to the accuracy of an interpretation of a test score. The term *valid test* is inappropriate. We consider the evidence supporting the creation of that test score as an accurate measure of student learning. By carefully creating OSI items and using these items in a test to obtain a test score fairly, we make a claim that the test score is a valid (accurate) measure of student learning.

That said, reliability comes into play. Reliability refers to the degree of random error represented in a set of test scores. We cannot have a validity interpretation of a test score, if reliability is low. Let us disregard how to compute reliability. Any measure of student learning should have a low degree of random error (thus high reliability). To ensure this valuable piece of validity evidence, OSIs MUST be well written and representative of the domain of knowledge and skills a test is supposed to represent. Longer tests tend to have less random error. Items of appropriate difficulty for the students tend to reduce random error. That is, items should not be too hard or too easy.

Students need to be informed about what they are about to learn. They much need to have a way to identify and learn what you are teaching. This content may be a lesson, unit, topic, course, curriculum, textbook, other written materials. Think of content as existing in domain that consists of knowledge and skills. Students learn the content in that domain. A test is a fair, unbiased sample from that domain to ensure high validity. If you guarantee the students have received adequate instruction and the test fairly represents this content, valid test score interpretations are achieved.

1.3. Content

We can categorize all content that is taught into four convenient categories.

1.3.1. *Facts*

Are true statements verifiable by all. The square root of nine is three. The area of a square or rectangle is the length of one side times the length of the adjacent side. Ankara is 445 kilometers from Istanbul. The opposite of East is West. Earth is a planet. Facts are notoriously over tested. We might say that facts are over taught. Focusing on facts does not leave room for more important types learning.

1.3.2. *Concepts*

A concept is an idea. For example, love, peace, fruit, car, money, television are some examples of concepts. Each concept has a definition, distinguishing characteristics, and examples. Thus, testing for a concept involves distinguishing among concepts, definitions of a concept, characteristics of the concept, or examples and non-examples of the concept.

1.3.3. *Principles*

Principles are relationships that are causal. Some principles are absolute (axiomatic), and some principles are probabilistic.

> *The first step in trauma injury is to ensure the airway is open. (Axiomatic)*
>
> *The density of air depends on its elevation. (Axiomatic)*
>
> *As temperature declines, at some point, water turns to ice. (Axiomatic)*
>
> *What is the chance of survival in an car accident if a passenger is wearing a seat belt. (Probabilistic)*
>
> *Which factors contribute to heart disease? (Probabilistic)*
>
> *Wheat tends grow optimally under what conditions? (Probabilistic)*

All formats presented previously can be useful for testing principles, but the testlet is the most highly recommended. Unfortunately, the testlet is difficult to design. However, many testlets can be designed to have interchanging values that provide more usability. That is, we can vary

values in a problem and create new problems and use a set of standardized questions as previously shown.

1.3.4. *Procedures*

A procedure is a set of mental or physical steps. OSIs are not suitable for measuring physical procedures. For mental procedures, we might ask a student to identify correct or incorrect sets of steps, or to identify a key feature of a procedure. Only the completion item is useful for measuring physical skills. The other OSI formats apply best to measuring knowledge.

We have plenty of understanding that facts are taught too much. Learning concepts is useful. Applying principles is more complex and very desirable in everyday life. Procedures are things we do every day and over time that have many steps. When you create a MC item, you will choose which of these four types of content will fit your purposes. Ultimately, we can to use facts, concepts, principles, and procedures in some combination that is complex. This leads us to mental complexity.

1.4. Mental Complexity

For every item, we assign a judgment of what type of mental complexity is required to choose the correct option. Of course, this is speculation, because every student has a different reaction to a MC item. The low-achieving student must use a higher degree of mental complexity in choosing a correct choice. The high-achieving student usually uses previous knowledge. Nonetheless, there is a premium of writing MC items with greater mental complexity because we want our students to use knowledge and skills in complex ways to solve problems, evaluate alternatives, decision-making, and thinking critically. Simply memorizing facts does not take us very far. Three types of mental complexity are briefly illustrated using the CMC format.

1.4.1. *Recall*

In Turkey, which river is the longest?

*A. Kizilirmak**

B. Euprhates

C. Tigris

This item may also be considered a trick item, because B and C are very long rivers but are shared by other countries. A is correct.

Items of this type are very easy to write and use. We have an abundance of recall items. Most educators admit that we tend to teach and test for recall instead of teaching for deeper and more complex types of student learning. Thus, recall items should be used sparingly.

1.4.2. *Understand (Comprehend)*

The focus here is a concept, which is an idea or mental picture of a group or class of objects formed by combining all their aspects. To measure a student's understanding of a concept we can ask them to identify the correct definition, the distinguishing characteristics, or examples of the concept.

OSI formats can also be designed to understand a principle or procedure.

Which of the following best defines the educational term assessment?

A. A student's test score

*B. A judgment based on a variety of valid information**

C. An evaluation of the student's mental, physical, and social conditions.

A is wrong because but many misuse this term. B is correct. C is too inclusive

Which of the following is an axiomatic principle?

*A. Longer tests tend to yield more higher test scores than shorter tests.**

B. A student test score is likely to be more accurate if the item difficulty matches the achievement level of the student.

C. The chances of correctly answer five CMC items correctly via random guess is very small.

A is correct because it is absolute. B and C are probabilistic therefore not axiomatic.

Which of the following influences the warming of the earth?

Mark A if true and B if false

1. The earth is closer to the sun.

2. Burning fossil fuels

3. Nuclear energy

4. Agriculture

5. Solar energy

6. Hydroelectric energy

1.4.3. *Application of knowledge and skills*

This category of mental complexity is most needed in modern education, because it requires students to use knowledge and skills in coordinated and complex ways. The most common examples are seen in testlets. In fact, the testlet designed to measure the application of knowledge and skills in complex ways. However, the truest form of application comes with a performance test where a checklist or rating scale is used and human judgment determines how well the student performs. Economies are gained by using OSIs for test items that measure application. Some examples of application testlet items are presented here is abbreviated form:

1. Reading. The student reads a passage and responds to three to 12 items probing various aspects of reading comprehension.
2. Mathematics. The often-used story problem initiates a testlet. As with reading, OSIs are used in a coordinated set.
3. History. A passage from a textbook is presented for student analysis. OSIs are presented as a set probe the students' ability to combine knowledge and skills to draw a conclusion, evaluate the merits of a decision, or extract a defensible analysis of the event.
4. Science. An experiment or a vignette introduces something the student was supposed to learn. The vignette might contain data, a chart, a graph, or a report. The OSIs probe who well the student understands and applies knowledge and skills.

1.5. Guidelines for Creating OSIs

Please explain the method, sample or study group, data collection tools, data collection process, and data analysis procedures in this section. This section should indicate the study's design, the sampling, the data collection tools, and the data analysis. Clarification is essential in this part.

In this section, some guidelines are highlighted to guide in the creating or evaluating OSIs. The basis for this section is a popular taxonomy has been published long ago and updated (Haladyna & Rodriguez, 2013). A list of guidelines appears on the internet and is widely shared and used. As a service to readers, poorly written items will be illustrated here as instruction for what not to do. These are really bad items.

Opinion Items. Which country offers the best kebabs? It might be factual, but it looks like an opinion.

Trick Items. In what country, do Panama hats originate? The correct answer is Chile. If one option is Panama, the student is tempted to choose that option.

Format Items Vertically, not Horizontally.

The Ankara Central Station represents which school of architecture?

A. Classical

B. Ottoman

*C. Modernism**

This is the clearest presentation a CMC item. However, in the interest of saving space, some test designers like to place option in the same line.

The Ankara Central Stations represents which school of architecture?

*A. Classical B. Ottoman C. Modernism**

This horizontal formatting may be confusing to some students.

Edit and Proof Items. All items should be grammatically correct and proofed. Common errors in sentence construction should be avoided. If an item is not well edited and proofed, it leaves a bad impression with the student. Also, lacking editing, the syntax of the item might be clumsy and by that confuse the student.

Linguistic Complexity, Window Dressing, Length. The reading level of any test items should be suitable for the reading level of the class. For those whose first language is different than the language used in a test, the linguistic complexity of an item stem might challenge the student unfairly. I am reminded of a licensing test for police where item stems were very long and linguistically very complex. Much of the information in the stem was irrelevant (window dressing). These factors led to very low performance on the licensing test. Remember that each item has a scoring weight of one. We should attempt to make each item as brief as possible yet retain the content and mental complexity needed.

Avoid Negation in the Stem and the Options.

Which is not true of cardiopulmonary resuscitation (CPR)?

*A. Closed chest massage is as effective as open chest message.**

B. The success rate for out-of-hospital resuscitation may be as high as 30% to 60%.

C. The most common cause of sudden death is ischemic heart disease.

Put the main idea in the stem of the item, not the options. The stem usually has more words than the options. However, one item-writing fault is the unfocused stem.

Agriculture

A. is an important part of the Turkish economy.

B. is an important part of the Turkish economy.

C. shows a decline in avocado production in Turkey.

For this kind of item, options may wander all over the place and even might not be grammatically equivalent.

All choices should be plausible? In writing or evaluating distractors, as the content expert, you are best suited to decide if a distractor is plausible. If it is not plausible, even a student who has not learned will eliminate that distractor and improve the chance of a lucky correct guess. Another way to find out if a distractor is implausible is to ask your students!

Avoid options such as none-of-the-above, all-of-the-above, and I-don't know. Such options offer clues for the clever student.

Longest option is correct. The weary item writer may write question and a long correct answer and then make the other choices due to lack of effort. The longer choice is the correct one.

Avoid absolute words. In the choices, certain absolute words are seldom correct. These absolutes include absolutely, always, completely, outright, never, perfect, without exception and ultimate. Again, clever students will avoid choices with extreme.

Repeating a word or phrase in the both the stem and one choice. This is a clue that the repeating word or phrase is correct. If it is not the right answer, then we have a trick.

What is are Mediterranean avocados principally grown in Turkey?

A. Mediterranean coastal region

B. Southern region

C. Northern region

Pairing terms that presents a clue.

Which condiments are best on kebabs?

A. Salt and pepper

B. Sugar and spice

C. Salt and spice

D. Spice and pepper

If a student does not know the answer, the choices may offer a clue. Spice appears three times. Salt and pepper twice. Sugar once.

Ridiculous Choices. In a hurry to find a third or fourth option, you might insert a choice that no student will choose.

In growing avocados, what is the most important factor?

A. Adequate water

B. An ideal climate

C. Good luck

D. A green thumb

An item like this one has essentially two plausible options.

Format Options in Numerical Order and Observe Place Value. In a test, most students can be very anxious and feel stress, putting numbers of numerical order with clear place value helps the student.

What is the speed of sound?

A. 120 km/h

B. 1200 km/h

C. 400 km/h

D. 700 km/h

What is the speed of sound in kilometers per hour?

A. 120

B. 400

C. 700

D. 1200

1.6. Creating a Collection of Items for Future Testing

This activity is difficult and time-consuming. Honestly, it takes years to develop a useful collection. This collection will also be subject to review: keep, revise, discard.

We have at least three ways to create a useful collection: (1) Free available items, (2) cloned items, (3) creating you own items. All three methods have advantages and disadvantages.

1.6.1. *Free items*

Depending upon the subject matter and students taught, the worldwide WEB provides many sources of free items. These items are open source. You can obtain such items easily and incorporate them into your item collection judiciously. Each item MUST represent suitable

content and have a desirable mental complexity that is appropriate for your students. The problem is that such items lack the close connection with actual instruction. Nonetheless, the price is right. If you can obtain some items at no cost, that might help you develop new, similar items, as suggested in the next strategy.

1.6.2. *Cloned items*

If you find items copyrighted, one strategy is to take the general form of the item and create a model. This is briefly illustrated with an item obtained from the Worldwide WEB.

Painting a wall that measure 15 square meters. One pint of paint covers 7 to 9 square meters. A pint of paint costs 100 Lira.

How much paint should I buy?

What will it cost?

If I have to use two coats, how much paint should I buy?

If I have to use two coats, how much will it cost?

A painter charges _____ per hour. She estimates the job to take _____ hours.

The above example is actually an outline for a testlet. It shows that with an item that contains area and cost for a product, many useful items can be generated. Automated Item Generation (Gierl & Haladyna, 2013) has many examples of item models that will produce many items. The limitation is that the items may measure a narrow band of content that is taught.

1.6.3. *Item shells*

Long ago, when helping pharmacists write useful test items for their national pharmacy licensing test, we came upon an idea that still works today (Haladyna & Shindoll, 1989). The approach we found useful is to identify items that had the same syntactic structure and create a shell of the item. The shell consisted on the stem followed by a blank where the content was inserted. Here are some examples of item shells (Haladyna & Rodriguez, 2013, p. 145). These are very generic.

Which is the best definition_____? Which is an example of _____? What is the meaning of _____? What is like _____? What are the distinguishing characteristics of _____?

Which is the principle of _____? What is the cause/reason for _____? What is the relationship between _____ and _____? Which is an example of the application of this principle _____? What would happen if _____? Which is better/worse, higher/lower, nearer/farther, heavier/lighter, _____? What is the difference/similarity between _____ and _____? Which principle best applies _____? What is the best way to ____?

One problem with item shells is that items generated from shells get to be repetitious. So, the use of any specific item shell should be limited. Nonetheless, the item shell gets item writers started if they have "writers' block." Clearly, it speeds up the item-writing process.

1.6.4. *Creating items*

The old-fashion way to create items is simply to select which format to use and write the item. Teacher/instructor-made test items are notoriously bad item writers. This tendency is true because most teachers/instructors do not have adequate training or have not been exposed to the formats, guidelines, and techniques found in this article and in the references are the end of this article.

Writing your own items is tedious and time-consuming. As pointed out previously, we often refer to your collection of items as an item bank. So, writing and placing items in your bank yields benefits in the future, just like a savings account in a bank.

1.7. Evaluating Items

Once items are created for measuring student learning in a classroom or course in a university or professional school, evaluating items is challenging. For large-scale testing programs, we have very sophisticated methods for evaluating test items (Haladyna, 2015; Haladyna & Rodriguez, 2021). These methods are inappropriate for student testing in the classroom.

In the classroom or in a course of study, how students respond to items is the best way to evaluate each item. A review of any summative test should reveal if items are working as intended. High-achieving students should choose correctly, and low-achieving students should choose incorrectly. If all students choose correctly, teaching has been effective and student learning has also been effective. If an item has a low degree of correct choice (less than 50% for a CMC item), we have a problem. Here are some questions that should help you evaluate whether your students are being given fair treatment in measuring what they have learned.

1. Is the item irrelevant regarding content?

2. Is the item flawed? Review the guidelines for writing items.

3. Does the item have two correct choices? This can happen.

4. Does the item have no correct choices? This can happen.

5. Was the content taught? Testing students on content not taught is not fair.

6. Did most of students dismiss what was taught? Students have to accept responsibility for a lack of study.

As we evaluate our test, we also evaluate our teaching. Honest discourse with students following the administration of a summative test, a meeting with students to go over test results reveals answers to the many questions just posed. Also, a chance for students to discuss what they learned and have not learned can be a valuable learning experience. It also helps you (the teacher/instructor) improve the quality of your collection of test items for future use.

1.8. Closing

The advice offered in this article is intended to guide you and your students toward a positive experience when it is time to measure what students have learned and help them continue on the path to future learning. Having a collection of useful test items is a start. Using these items in formative and summative ways is important as we guide each student to a successful end of their brief educational experience.

Declaration of Conflicting Interests and Ethics

The authors declare no conflict of interest. This research study complies with research publishing ethics. The scientific and legal responsibility for manuscripts published in IJATE belongs to the author.

Orcid

Thomas Haladyna https://orcid.org/0000-0003-3761-6979

REFERENCES

Gierl, M., & Haladyna, T.M. (Eds.). (2013). *Automatic item generation: Theory and practice.* Routledge.

Haladyna, T.M. (Submitted for publication). *How much of threat to validity is random guessing?*

Haladyna, T.M. (2015). Item analysis for selected-response test items. In Lane, S., Raymond, M.R., & Haladyna, T.M. (Eds.). *Handbook of test development* (pp. 392-409). Routledge.

Haladyna, T.M., & Rodriguez, M.R. (2021). Using full-information item analysis to evaluate multiple-choice distractors. *Educational Assessment, 26*(3), 198-211. https://doi.org/10.1080/10627197.2021.1946390

Haladyna, T.M., & Shindoll, L.R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation in the Health Professions, 12*(1) 97-106. https://doi.org/10.1177/016327878901200106

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

Haladyna, T.M., Raymond, M.R., & Stevens, C. (2019). Are multiple-choice items too fat? *Educational Assessment, 32*(4), 350-364. https://doi.org/10.1080/08957347.2019.1660348

Lane, S., Raymond, M., & Haladyna, T.M. (Eds.) (2015). *Handbook of test development* (2nd ed.). Routledge.

Rodriguez, M.C. (2016). Selected-response item development. In Lane, S., Raymond, M.R., & Haladyna, T.M. (Eds.). *Handbook of test development* (pp. 259-273). Routledge.