

## Assessment Tips and Best Practices for OHSU YourMD Educational Leaders

Last Updated: February 2025

### Contacts:

Tracy Bumsted, MD, MPH  
 Associate Dean, UME  
[bumstedt@ohsu.edu](mailto:bumstedt@ohsu.edu)

Will Fruhwirth, MBA  
 UME Assessment Program Manager  
[fruhwirt@ohsu.edu](mailto:fruhwirt@ohsu.edu)  
 503-494-8017

Emily Larson  
 Administrative Manager, UME Teaching Services Office (TSO)  
[larsonem@ohsu.edu](mailto:larsonem@ohsu.edu)  
 503-494-7121

UME Teaching Services Office (TSO)  
[tso@ohsu.edu](mailto:tso@ohsu.edu)  
 503-494-8428

### Reminders for FoM UME leaders as of February 2025:

FoM Assessment Component	Primary Responsible Person(s) for Creating and Submitting Test Questions for YourMD Assessments
<b>1</b> (weekly formative quizzes)	<b>All Block Directors</b> +/- Thread Directors depending on weekly content in Block
<b>2A</b> (weekly Clinical + Health Systems Science skills)	2A – Clinical Skills Assessments (CSA) - Clinical Thread Directors ( <b>Pete Sullivan, Cliff Coleman, Karen Anstey</b> ) 2A – HSS/Informatics Skills Assessments – HSS Thread Director <b>Karen Anstey + Gretchen Scholl</b> , Educational Informaticist
<b>2B</b> (weekly Basic Science skills)	2B – Basic Science Thread Directors ( <b>Sylvia Nelson, Peter Mayinger, Richelle Malott, Erin Bonura</b> )
<b>3</b> (ExamSoft Final Exam, 1/block)	<b>All Block Directors</b> +/- Thread Directors and faculty instructors depending on content
<b>4</b> (NBME Final Exam, 1/block)	<b>All Block Directors</b> +/- Thread Directors depending on content
<b>5A</b> (final Clinical + Health Systems Science skills)	5A – Objective Structured Clinical Exams (OSCEs) - Clinical Thread Directors ( <b>Pete Sullivan, Cliff Coleman, Karen Anstey</b> ) 5A – HSS/Informatics Skills Assessments – HSS Thread Director <b>Karen Anstey + Gretchen Scholl</b> , Educational Informaticist
<b>5B</b> (final Basic Science skills)	5B – Basic Science Thread Directors ( <b>Sylvia Nelson, Peter Mayinger, Richelle Malott, Erin Bonura</b> )

# Writing Strong Multiple Choice Questions

Component 1 weekly quizzes are formative and consist of multiple choice questions (MCQs) that test the **main concepts** covered in class or in lab and are clearly connected to the Block **objectives**. Component 3 final exams are summative, graded, and are cumulative.

- Questions can reference **central** points in required reading as linked to objectives
- Each question must include an **annotated answer** or "**rationale**" explaining why the correct answer is the only correct answer – it should "teach."

MCQs should be as straightforward as possible. Aim to make the question stem concise and content rich so that the answer choices are clear.

The following steps may be helpful as you design your MCQs:

**Step 1 - The Question:** What is the main idea you want to assess?

- Ask the question as directly and succinctly as you can.
- Use positively stated question stems (avoid - "this is not a diagnosis that doesn't involve which of the following?")
- No True/False questions
- The test taker should be able to cover up all answer choices and still be able to identify the correct answer to the question after reading the stem.

**Step 2 - Answer choices:** What is the 1 correct answer?

- Limit answer choices to A-D (including the correct answer)
- One answer is correct and the remaining three should be clearly incorrect (i.e., no answers that could be argued as plausible. or "somewhat correct")
- No answers with "All of the above," "A + C" or "None of the above"
- No True or False questions (e.g., "which of the following are false?")

## EXAMPLES

### WEAK QUESTION 🙅

*The nerve(s) that travels from the neck to the diaphragm on either side of the pericardial sac is...*

- Cardiac splanchnic*
- Phrenic*
- Vagus*
- A and C*

Answer B.

The question is phrased in a sentence fragment that doesn't ask a question. The answer relies heavily on simple memorization. No objectives or rationale are included and it has a confusing answer choice (d.)

# Writing Strong Multiple Choice Questions

## STRONG QUESTION

*A 38-year-old man is admitted to the hospital with sharp chest pain just behind his sternum. He is diagnosed with acute pericarditis (infection of the pericardial sac). The nerve that is transmitting the painful sensation is which of the following?*

- a. cardiac splanchnic
- b. intercostal
- c. phrenic
- d. vagus

**Rationale:** *The quality of the pain suggests somatic afferent sensation from a somatic tissue. The parietal and fibrous pericardium are somatic tissues, and they are innervated by the phrenic nerve, which travels from the neck to the diaphragm along either side of the pericardial sac. The intercostal nerves would carry pain from parietal pleura on the costal surface of the lungs. The vagus and cardiac splanchnic nerves are visceral nerves.*

**Objective:** *Describe the motor innervation and sensory fibers of the diaphragm*

This is a strong question because it asks the learner to apply what they know. The rationale provides relevant information about why the other answer choices are incorrect and explains why the correct answer is the only possible choice. It is also linked to one of the class objectives.

### More tips:

- Create test questions using full/complete sentence structure with a question mark.
- If there is a patient case, describe the case (a.k.a., vignette) and then ask a question.
- Drug names are not capitalized **unless** it's a brand name (e.g., Viagra, Flonase vs. acetaminophen, codeine).
- Indicate the correct answer by underlining or highlighting text.
- Please make sure that there are only 4 answer choices - one correct answer and three distractors.
- Remember that many questions should end with "which of the following?", as in, "The tapeworm that humans would most likely contract from eating undercooked beef is which of the following?" The reason for creating questions that end with "which of the following?" is not random or idiosyncratic. The literature is pretty clear that students get quickly used to a pattern and can anticipate that the important stuff/cognitive load for each question is prior to the very end. Having a mix of questions just makes things harder and not in the way we want (mastery of content.) It's similar to always having the same pattern of a SOAP note or oral presentation.

## Procedures for Creating an ExamSoft Assessment

*\*Block directors have the option to create and edit exams/quizzes directly in ExamSoft and skip the instructions below. If you would like to do this, contact TSO for access and instructions.*

At the beginning of your block, TSO will download updated versions of each week's quiz from the previous year and save them to the appropriate Assessment folder in the UME Shared OneDrive.

- These quiz downloads contain edits and refinements to quiz questions made in previous years and represent the most up-to-date version.
- Any new questions received from thread directors and other instructors will also be saved in the specific weekly folder.
- You can make any edits directly in the draft quiz document. You may edit existing questions, add new questions, or remove any questions that you don't want to use. ALWAYS use track changes.
- When adding a new question, please be sure to add a question title, correct answer, rationale, objective, and any relevant threads you would like the question tagged with.
- Please include a helper note next to the question number. (*Example: Question #10: Edited, or Question #12: New. You can also add a note that says "Delete" if you would like to remove a question.*)
- Please leave all ExamSoft data (especially **Item ID #**) in the document. TSO uses this information to edit the quiz in ExamSoft.
- If you would like any additional quizzes exported from ExamSoft from prior years, please ask TSO, and they will provide that as well.

**Your edited draft quiz is due to TSO by Wednesday morning each week to ensure on-time publication for the Friday morning quiz.**

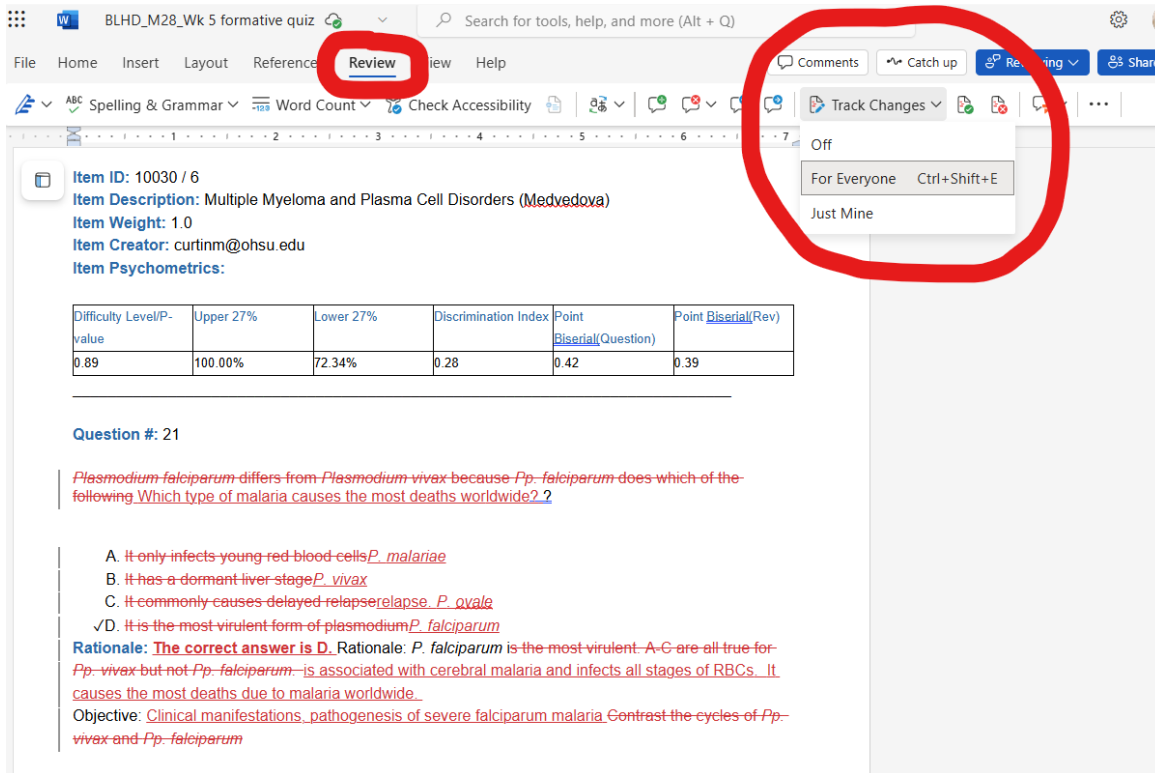
Please **ALWAYS USE TRACK CHANGES** when editing the quiz document.

- Using **Track Changes** helps ensure that TSO will see all your edits and saves A LOT of time for TSO staff.
- If you do not use **Track Changes**, your edits may not appear in the published quiz.
- See more details and screenshots of **Track Changes** on the next page.

## To turn on Track Changes:

- Click the **Review** tab at the top of the screen.
- Click **Track Changes** near the right side of the ribbon.
- Select **For Everyone**.

Now, any edits you make to the document will appear alongside the original text of the document.



The screenshot shows the Microsoft Word interface with the **Review** tab selected. The **Track Changes** button is circled in red. A dropdown menu is open, showing the following options: **Off**, **For Everyone** (with the keyboard shortcut **Ctrl+Shift+E**), and **Just Mine**. The document content includes item metadata and a table of psychometric data.

Difficulty Level/P-value	Upper 27%	Lower 27%	Discrimination Index	Point Biserial(Question)	Point Biserial(Rev)
0.89	100.00%	72.34%	0.28	0.42	0.39

**Question #: 21**

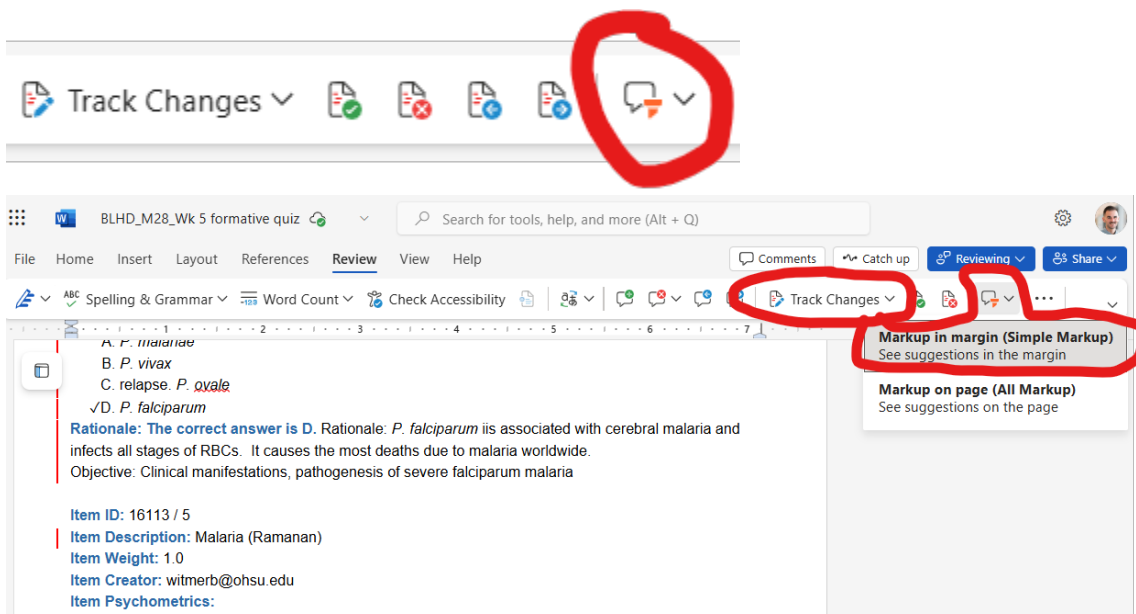
*Plasmodium falciparum* differs from *Plasmodium vivax* because *P. falciparum* does which of the following Which type of malaria causes the most deaths worldwide? ?

- A. It only infects young red blood cells *P. malariae*
- B. It has a dormant liver stage *P. vivax*
- C. It commonly causes delayed relapse *P. ovale*
- ✓ D. It is the most virulent form of plasmodium *P. falciparum*

**Rationale:** The correct answer is D. Rationale: *P. falciparum* is the most virulent. A-C are all true for *P. vivax* but not *P. falciparum*—is associated with cerebral malaria and infects all stages of RBCs. It causes the most deaths due to malaria worldwide.

**Objective:** Clinical manifestations, pathogenesis of severe falciparum malaria. Contrast the cycles of *P. vivax* and *P. falciparum*.

This might look distracting. Don't worry – you can hide the Track Changes mark-up, while leaving it visible for TSO!



The screenshot shows the Microsoft Word interface with the **Review** tab selected. The **Track Changes** button is circled in red. A dropdown menu is open, showing the following options: **Markup in margin (Simple Markup)** (with the subtext "See suggestions in the margin") and **Markup on page (All Markup)** (with the subtext "See suggestions on the page"). The document content includes item metadata and a list of multiple-choice options.

- A. *P. malariae*
- B. *P. vivax*
- C. relapse *P. ovale*
- ✓ D. *P. falciparum*

**Rationale:** The correct answer is D. Rationale: *P. falciparum* is associated with cerebral malaria and infects all stages of RBCs. It causes the most deaths due to malaria worldwide.

**Objective:** Clinical manifestations, pathogenesis of severe falciparum malaria

**Item ID:** 16113 / 5

**Item Description:** Malaria (Ramanan)

**Item Weight:** 1.0

**Item Creator:** witmerb@ohsu.edu

**Item Psychometrics:**

# Test Item Difficulty, Discrimination Index, and Point Biserial

Explained by Tom Boudrot, Ed.D., former Director of OHSU Teaching & Learning Center

The **Point-biserial** and **discrimination index** are almost identical (simply different calculations that achieve the same purpose – not sure why ExamSoft displays both) and compare students in the top 27% in overall test performance from the bottom 27% (overall as well) on a **specific test item**. The discrimination scores are useful measures of item **quality** whenever the purpose of a test is to produce a spread of scores, reflecting differences in student achievement. I use the point-biserial and discrimination index interchangeably.

I've taken a few slides from a workshop I have done often that shows two important metrics:


**P-value** (item difficulty) and **IDis** (item discrimination or point-biserial). They're both pretty basic and easy to compute. The item discrimination slide shows the computation when five students in the upper group and 2 students in the lower group got a specific exam item correct yielding a discrimination index of .50 (a good discriminator).

I also included two slides that show the basic ranges of each value. For test items that aren't written by professional test developers. I give more wiggle room for item discrimination and consider anything above .20 to be "in the ball park" and anything above .30 to be pretty good.


### Item Difficulty

**p-value**

42 students answered the item



8 got it correct


$$\frac{\text{\# Who Got the Item Correct}}{\text{\# of Students who Answered the Item}} = \frac{8}{42} = .19$$


### Item Difficulty

**p-value range**


The higher the value, the *easier* the item.

- **Above 0.90** -- too easy; review for question's purpose (warm up? fundamental?)
- **Below 0.20** -- too difficult; review for confusing language, remove from subsequent exams, and/or identify as area for re-instruction.




### Item Discrimination

**point-biserial correlation**



$$\frac{(\text{\# Upper Group Correct}) - (\text{\# Lower Group Correct})}{\text{Number of Students in the Upper Group}} = \frac{5 - 2}{6} = .50$$

Image Sources: [http://www.allaroundtrainschool.com/blogstockphoto\\_Happy\\_Group\\_Of\\_Friends\\_2134473.jpg](http://www.allaroundtrainschool.com/blogstockphoto_Happy_Group_Of_Friends_2134473.jpg)  
[http://igsupermarthe.com/secondary/wp-content/uploads/2009/06/sad\\_group2.jpg](http://igsupermarthe.com/secondary/wp-content/uploads/2009/06/sad_group2.jpg)




### Item Discrimination

**IDis = (Upper Group % Correct) – (Lower Group % Correct)**

Negative ID	Unacceptable – check for item error
0% - 24%	Usually unacceptable
25% - 39%	Good item
40% - 100%	Excellent item

Adapted from University of Wisconsin Oshkosh:  
<http://www.uwosh.edu/testing/facultyinfo/itemdiscrimone.php>

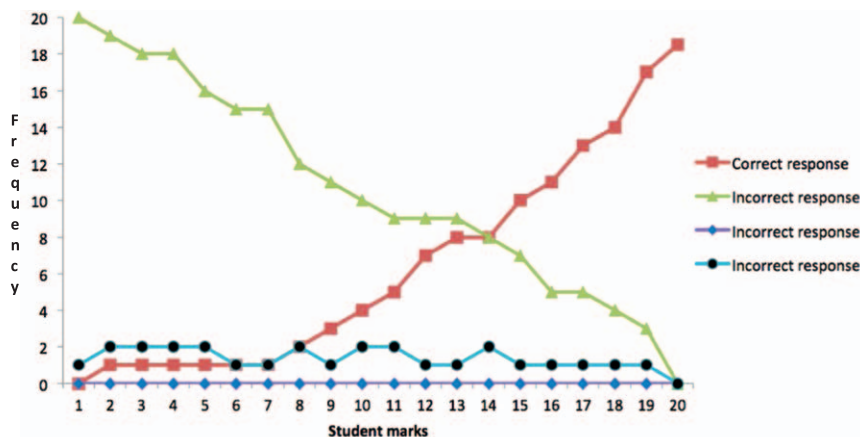


## Postexamination Analysis: A Means of Improving the Exam Cycle

Mohsen Tavakol, PhD, MEd, assistant professor of psychometrics, and Reg Dennick, PhD, professor of medical education, University of Nottingham, Medical Education Unit

The exam cycle involves all steps needed to create, run, and evaluate an exam. It includes deciding on learning outcomes, writing questions, setting assessment standards, and providing feedback to learners and assessors. Postexam analysis is conducted by psychometricians after the assessment is administered and scored. Since the exam cycle is a recursive process, postexam analysis continues until medical educators obtain reliable and valid evidence that the exam is accurately measuring student performance. Post exam analysis can determine, for example, if most students correctly answering a question means that the students have grasped the material or that the question is too easy. Postexam analysis is used for all levels of medical education and all types of objective tests including OSCEs. Analysis is performed by applying well-established psychometric procedures. The purpose of this AM Last Page is to briefly describe two of the key methods.

**Item analysis.** Item analysis can improve the quality of a test by detecting questions that are vague, ambiguous, too easy, or too difficult. One type of item analysis—item difficulty—refers to the percentage of a cohort of students who respond to a question correctly (also called the P value, not to be confused with the P value related to statistical hypothesis testing). The higher the P value, the easier the question. If all students answer a question correctly, the P value is 1.0 (too easy), and if all students answer the question incorrectly, the P value is 0.0 (too difficult). Questions that are too easy or too difficult should be removed from assessments as they do not discriminate between high and low performers and hence have poor item discrimination ( $d$ ). Item discrimination refers to the capacity of a question to discriminate among students who perform well and those who do not. As P value increases from 0 to 0.5, the  $d$  value increases; however, as P value increases from 0.5 to 1.0, the  $d$  value decreases.



**Trace lines.** Trace lines are defined as a display of the frequency distribution for selecting options in a multiple-choice question from a cohort of students. These lines help educators make judgments about the plausibility of options in multiple-choice tests. For example, the flat traces in Figure 1 show that these two incorrect options were not plausible. They were too easily eliminated and thus did not discriminate between students.

### Postexam analysis can improve the exam cycle by:

- Providing standard setters with valuable information for understanding a borderline student before judging a question,
- Enabling students to identify their weaknesses and strengths in order to improve their learning,<sup>1,2</sup>
- Providing validity and reliability evidence for test scores,
- Detecting aberrant or ambiguous questions that can be modified or discarded,
- Developing effective multiple-choice questions by removing implausible options,
- Developing fair assessment questions, especially for students of different groups (e.g., gender, race/ethnicity),
- Minimizing identified sources of error to improve the reliability of scores (e.g., examiner bias in OSCEs),
- Constructing “testlets” to improve the quality of assessment,
- Establishing an item bank that can be used in national assessment,
- Developing multiple forms of an assessment and equating them,
- Developing a fair and legally defensible pass mark, and
- Adjusting student marks to flawed items.

**Testlets**—or item bundles—are groups of questions that are highly correlated to one another and relate to a common construct. Testlets can reduce exam time and improve validity and reliability.

#### References

1. McDonald ME. Guide to Assessing Learning Outcomes, 2nd ed. Sudbury, Mass: Jones & Bartlett Learning; 2014.
2. Tavakol M, Dennick R. Post examination analysis of objective tests. Med Teach. 2011;33:447–458.

**Author contact:** mohsen.tavakol@nottingham.ac.uk.

OSCE indicates objective structured clinical exam.