TEXAS TECH UNIVERSITY HEALTH SCIENCES CENTER EL PASO
Paul L. Foster School of Medicine

# Medical Education Program Policy

| | | | | | |
|---|---|---|---|---|---|
| **Policy Name:** | **Summative Test Item Performance Policy** | | | | |
| **Policy Domain:** | Assessment | | **Refers to LCME Element(s):** | | |
| **Approval Authority:** | Curriculum and Educational Policy Committee | **Adopted:** | 2/1/2016 | **Date Last Reviewed:** | 2/1/2016 |
| **Responsible Executive:** | Associate Dean for Medical Education | **Date Last Revised:** | | | |
| **Responsible Office:** | Office of Medical Education | **Contact:** | robin.dankovich@ttuhsc.edu | | |

1. **Policy Statement:** Individual test item quality on pre-clerkship multiple choice question-based summative exams must maintain a level appropriate in assessing student understanding. This policy establishes the criteria for test bank items with standards that assess the reliability and validity of items beginning with the Academic Year 2016-17.
2. **Reason for Policy:** PLFSOM administers NBME style exams to pre-clerkship students as a means of assessing the students' knowledge base. While we recognize the importance of subject mastery, these exams are intended to provide a reliable and valid means of assessing the overall knowledge base of the student. The quality of individual test items on a test determines the reliability and validity of that test. With this in mind, this policy sets the standards by which test items will be kept in the test bank.
3. **Who Should Read this Policy**:
    - Pre-clerkship Phase (Year 1 and Year 2) Course Directors and Course Faculty
4. **Resources**: Office of Medical Education Annual Evaluation Report
5. **Definitions:**
    - "Item difficulty" – calculated as percentage of the class getting item correct.
    - "Item discrimination" – calculated as the percentage of students in the upper quartile who get the correct answer minus the percentage of students in the lower quartile who get the correct answer
6. **The Policy:**

    **Reporting and Monitoring:**

    - *Data indicating test item quality will be published as part of the Office of Medical Education Annual Report for CEPC review.*
    - *The Assistant Dean for Medical Education for Basic Science Instruction and the Year 1-2 Committee will review the data resulting from the application of this*

*policy after each SPM unit (as part of the unit debriefing). The CEPC will review the data in aggregate on an annual basis – or as deemed necessary by the Assistant Dean for Medical Education for Basic Science Instruction based on the outcome of the unit reviews.*

- *Benchmark data established AY 2016-17, the initial implementation period of this policy*

**Items requiring action:** *Test items that do not perform within the quality guidelines will be removed from the test item pool, pending either improvement or replacement.*

- Difficulty
  - For any item with a difficulty of .2 or less, the item will be removed from the test and from the pool until improved (see below).
  - For any item with a difficulty of .9 or above, no changes to the test are required. The item is removed from the pool until it is made more difficult.
- Discrimination
  - Items with discrimination scores less than .1, item is removed from the pool until improved.
- Foil Quality
  - If 50% or more of the foils are not selected, the item is removed from the pool until improved.
  - Items that fall within the quality guidelines will be included in grade calculations. Figure 1 presents the flow of decision points about item actions.

**Item Remediation Process:** *When an item is removed from the test bank/item pool, the responsible faculty member shall have the option of permanently archiving the question or improving the question. If the item is archived, it will be tagged as unusable so that it may not be used again without improvement.*

*If the faculty chooses to improve the question, a team of at least 2 other faculty members shall review the question. The reviewers will be provided with the original item statistics and reason for revision.*

7. **Attachments:** The attached document entitled, "Summative Test Item Standards Policy" (as approved by the CEPC on February 1, 2016) is adopted as a Medical Education Program Policy.

# Summative Test Item Standards Policy

## Purpose:

PLFSOM administers NBME style exams to the M1 & M2 students as a means of assessing the students' knowledge base. While we recognize the importance of subject mastery, these exams are intended to provide a reliable and valid means of assessing the overall knowledge base of the student. The quality of individual test items on a test determines the reliability and validity of that test. With this in mind, this policy sets the standards by which test items will be kept in the test bank.

## Item Statistics used by this policy

**Item difficulty** – calculated as percentage of the class getting the item correct.

**Item discrimination** – calculated as the percentage of students in the upper quartile who get the correct answer minus the percentage of students in the lower quartile who get the correct answer.

## Items requiring action

Test items that do not perform within the quality guidelines will be removed from the test item pool, pending either improvement or replacement.

- Difficulty
  - For any item with a difficulty of .2 or less, the item will be removed from the test and from the pool until improved (see below).
  - For any item with a difficulty of .9 or above, no changes to the test are required. The item is removed from the pool until it is made more difficult.
- Discrimination
  - Items with discrimination scores less than .1, item is removed from the pool until improved.
- Foil Quality
  - If 50% or more of the foils are not selected, the item is removed from the pool until improved.

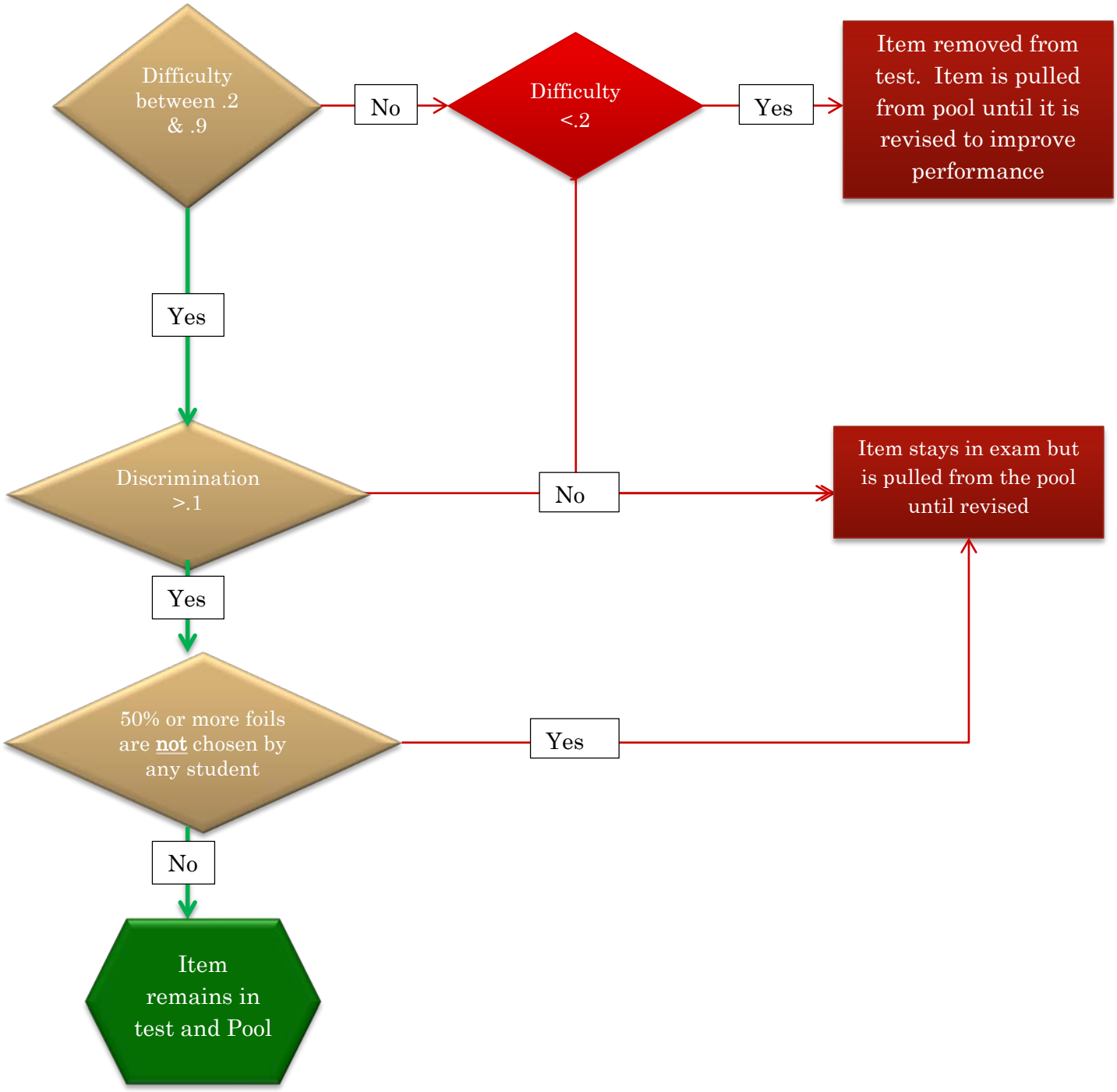
Items that fall within the quality guidelines will be included in grade calculations. Figure 1 presents the flow of decision points about item actions.

## Item Remediation Process

When an item is removed from the test bank/item pool, the responsible faculty member shall have the option of permanently archiving the question or improving the question. If the item is archived, it will be tagged as unusable so that it may not be used again without improvement.

If the faculty chooses to improve the question, a team of at least 2 other faculty members shall review the question. The reviewers will be provided with the original item statistics and reason for revision.

## Figure 1: Item Analysis Decision Flow



## Annotated Bibliography:


**Crystal Ramsay**, *Item Analysis*. Accessed at http://sites.psu.edu/itemanalysis/difficulty-2/ - provides a short tutorial on item statistics. Information used for this policy:

| % Correct | Item difficulty designation |
|---|---|
| 0 – 20 | Very difficult |
| 21 – 60 | Difficult |
| 61 – 90 | Moderately difficult |
| 91 – 100 | Easy |

*"Very easy or very difficult items are not good discriminators…. It is typically recommended that item discrimination be at least .20."*

**Office of Educational Assessment**, *Understanding Item Analysis Reports*. Accessed at https://www.washington.edu/oea/services/scanning_scoring/scoring/item_analysis.html. Information used for this policy:

*Ideal difficulty levels for multiple-choice items in terms of discrimination potential are:*

| Format | Ideal Difficulty |
|---|---|
| Five-response multiple-choice | 70 |
| Four-response multiple-choice | 74 |
| Three-response multiple-choice | 77 |
| True-false (two-response multiple-choice) | 85 |

*(from Lord, F.M. "The Relationship of the Reliability of Multiple-Choice Test to the Distribution of Item Difficulties," Psychometrika, 1952, 18, 181-194.)*

**Scoring Office, Michigan State University**, Item Analysis Guidelines. Accessed at https://www.msu.edu/dept/soweb/itanhand.html.

*… If possible, items should have indices of difficulty no less than 20 and no greater than 80. It is desirable to have most items in the 30 to 50 range of difficulty. Very hard or very easy items contribute little to the discriminating power of a test.*

**Kehoe, Jerard** (1995). Basic item analysis for multiple-choice tests. *Practical Assessment, Research & Evaluation*, 4(10). Retrieved October 13, 2015 from http://PAREonline.net/getvn.asp?v=4&n=10

*The proportion of students answering an item correctly also affects its discrimination power. This point may be summarized by saying that items answered correctly (or incorrectly) by a large proportion of examinees (more than 85%) have markedly reduced power to discriminate. On a good test, most items will be answered correctly by 30% to 80% of the examinees…. Distractors that are not chosen by any examinees should be replaced or eliminated. They are not contributing to the test's ability to discriminate the good students from the poor students. … Items that virtually everyone gets right are useless for discriminating among students and should be replaced by more difficult items. …*

French, Christine (2001). A Review of Classical Methods of Item Analysis. Annual Meeting of the Southwest Educational Research Association (New Orleans, LA, February 1-3, 2001). Accessed at http://files.eric.ed.gov/fulltext/ED450152.pdf.

*… A high index of item discrimination (d > .40) will always preferred over a lower index of discrimination (Ebel & Frisbie, 1986). …The item discrimination index is equal to the number of students in the upper scoring group, U, minus the number of students in the lower scoring group, L, who get the correct answer on a certain question. The difference is then divided by the total number of students in each group (Cohen, Swerdlick, & Phillips, 1996).*

*However, there is a general rule about the preference level for an item discrimination index. Anastasi and Urbina (1997) suggested a level above or as close to 50% as possible. Others have laid out a guideline of all the possible discrimination index values and their evaluation. Ebel and Frisbie (1986) suggested that item discrimination indices greater than .40 are very good items, those between .30 and .39 are good but there is some room for revision, those between .20 and .29 are borderline and are in need of improvement, and those below .19 should be eliminated or undergo much improvement (p. 234).*

**McCowan** , Richard N and Sheila C. McCowan, 1999. *Item Analysis for Criterion- Referenced Tests.* Buffalo, New York 14207-2407.

*Table 9*
*Optimal Difficulty Levels for Items with Different Options*
*(for tests with 100 items)*

| *Optimal Difficulty Level* | *Number of Options* |
|---|---|
| *2* | *.75* |
| *3* | *.67* |
| *4* | *.63* |
| *5* | *.60* |