CoAEMSP
Accreditation is *credible* education

Committee on Accreditation
of Educational Programs for the
Emergency Medical Services Professions

CoAEMSP

Item Analysis Q & A

The following questions were posed during the webinar Item Analysis for Written Exams *on June 30, 2015.*

Cut Score

1. **What is a cut score?**

   The minimum passing score on an exam or evaluation.

2. **How do you determine the cut score?**

   This is a somewhat complicated question. Please reference the *Foundations of Education: An EMS Approach 2nd Ed.*, pages 320 – 322.

3. **Where does the CoAEMSP stand on cut scores and their use to grade exams?**

   The CoAEMSP believes there is merit in evaluating an appropriate cut score for each exam; however, recognizes that this may not be realistically achieved in many institutions. Using external exams (i.e., NREMT or the State exam) and looking at how students perform on the program's exams may be a good indicator of whether or not the program's exams need to be "ratchet up" or changed in any way.

4. **What reading level should we be aiming for when writing exam items?**

   Items should be written at the reading level as determined in the functional job analysis. The job analysis expectation is that EMTs will read at the 10th grade level, and paramedics will read at the 13th grade and above level. The NREMT currently writes the EMT exam at the 8th grade level, and the paramedic exam at the 10th grade level.

5. **If an exam is given that could result in removing a student from the program, several students miss the cut score by a few points, and a few items are found that do not perform like predicted, should those items be thrown out?**

   Possibly. One should consider the items in light of current analysis (RPBI & difficulty level), previous performance of the item on exams, number of students in the pool answering the item, the item content, etc. Items could be tossed or two answers taken. Such an evaluation should be done blindly—evaluating items without knowing the outcome for particular students.

Validity and Reliability

6. **As a new program director, how many cohorts should I consider before I make changes to testing questions?**

   It depends on the number of students who have seen and responded to the item. A single class of 40 might be a decision maker; or 4 cohorts of 6 students might not be enough. A minimum of 40 students would be a good indicator unless you see the problem with the item despite the data. (The item can be problematic content wise without a problem in the data.) If the item is a negative RPBI, the program should change it or re-work it or toss it for the next exam. One must also be careful of giving the same exams without changing a percentage of the items for a next class.

7. **What do you think about validated tests through FISDAP and EMSTesting?**

   The CoAEMSP does not endorse any specific product and encourages programs to inquire with the vendors to determine if a product meets their needs.

8. **Our program uses Platinum Education-EMS Testing as our high stakes exams. Is that considered valid and reliable by the CoAEMSP?**

   The CoAEMSP encourages programs that use commercially available valid and reliable exams to inquire with the vendor to determine the degree of validity and reliability offered. If a program chooses to use a commercial product, it is crucial to measure students at an individual level, not just against other students in the nation.

Point Biserial

9. **How do you calculate the biserial score? I have a spreadsheet; however, it only gives one score for the question, not one for each answer.**

   If there is a single statistic reported in a report, it is most likely being reported for the keyed answer for the item. Without seeing the actual report it is difficult to be sure on the answer for this question.

10. **What software (free or for a fee) is available for analysis?**

    Scantron makes a program, Par Score. You are encouraged to conduct an internet search for exam item analysis software. Most of the robust learning management systems (LMS) have the capability to complete item analysis. Using an LMS may require the exams be setup in the LMS and given by computer.

11. **How do you calculate the RPBI (point biserial correlation)?**

    This is a complicated answer and beyond the scope of these educators; it is a complex mathematical formula best accomplished by a computer.

12. **What is KR20 and KR21?**

    These are reliability indices – Kuder-Richardson. For more information, reference the *Foundations of Education: An EMS Approach*, page 291.

13. Would it be an accurate statement to say: "Point Biserial measures the amount of correlation between difficulty and discrimination?"

The point Biserial measures the correlation between high performance on the exam and high performance on that particular item. It measures the effectiveness of the item.

Difficulty index and item discrimination measure different aspects of item performance. There is not a direct correlation between difficulty index and item discrimination, though you may see valid and reliable items that have a high degree of difficulty also effectively discriminate high versus low performance on an examination item. RPBI discriminates how those with the highest scores on the exam perform on each item in comparison to those students with the lowest overall scores. RPBI scores must be positive (0 - +1), with higher scores demonstrating more strongly the effectiveness of an item to discriminate top performers. Items with a negative RPBI are discriminating against the top performers and need to be critically analyzed for item effectiveness. There are many reasons why an item might have a negative RPBI, including, an overly tricky item, there is more than one correct answer, teaching was confusing or inaccurate, or the item may have been miskeyed. These are the reasons why it is so important to analyze examination item performance.

Discrimination Index

14. What is Deb's down and dirty way to evaluate the discrimination index?

This is one method to help identify trends on items to evaluate discrimination.

1) Divide the class into 3 groups based on each student's score on the exam you are evaluating.

2) Put the exams in 3 stacks with the high performers, medium performers, and low performers with approximately the same number in each group. If there are scores that are the same they should go together in a group.

3) Take item # 1. If 90 – 100% of the students got this item correct, move on to item #2 because items that are easy (high difficulty index) do not discriminate.

   If item # 1 has a difficulty level of 85% or below, look and see if the high performer group all got the item correct. If they did, that is a good sign that the high performers did well on that item, which means from the perspective of the high performers, the item is effective. If some of the high performers got this item wrong, then that is a sign that the item has a reverse discrimination and likely NOT an effective item. This item may need to be thrown out, especially if there are several high performers that missed this item.

4) Continue looking at item #1. Look and see how the medium performers did on the item. If all of the medium performers got it correct then it may or may not be that the item is doing what it is supposed to do and may or may not be effective. If some students in the medium performer group missed it and some got it correct, then that also means it may or may not be effective. The middle group may not tell you much.

5) Continue looking at item #1. Look and see how the low performers did on the item. If all students in the low performing group got the item WRONG, then the item is doing what it's supposed to do—discriminate against the low performers. This would be evidence that the item is effective.

6) Continue through all the items this way.

15. When the term "negative" discrimination is used, is this referring to the key or each distractor as well?

Negative discrimination is for that item—the keyed correct answer will be negative. The only important RPBI information is the one on the correct answer.

A computer-analyzed exam will have an RPBI for each item. Focus on the RPBI calculated for the key (correct answer). This number will range from -1.0 to +1.0, with 0 representing no discrimination. You want to see positive numbers and the closer to +1 (it is common to see solid questions in the +0.25 to +0.6 range), the greater the degree of discrimination between the students with the highest versus the lowest scores on the overall examination. Negative results for the keyed answer demonstrates that the students with the strongest overall performance are getting this particular item incorrect, and those students with the lowest scores are getting the item correct. Ask yourself, why does this discrepancy exist?

16. Should the *discrimination index* only be run on the "high" performers then or on the entire class?

The discrimination index or RPBI is run on every item on the exam, and uses the “graded” exam to calculate the index. It should be done on each item.

The discrimination index or RPBI is valid only when considering the responses of all students to each item contained on the examination. This calculation is made by using data for each individual item and the composite scores for every student completing the exam, comparing how each item performed for those with the highest scores versus those with the lowest scores.

17. Do "high performers" change on every exam?

Probably, but not necessarily. The exam is graded and the high performers on that exam are what are used to calculate the discrimination index. It is really the “high performers” on that specific exam. Students may have varied areas of strength and weakness, and this may be reflected in specific unit exams or content areas.

18. Does "high performer" mean above a particular score on the test being scored at the moment?

High performer only means higher than others on that exam. So the students are ranked based on their score and those who are closer to the top of the scores are the higher performers.
It may be helpful to think of this similar to a bell-shaped curve.

19. How does a missed question by a high performer affect the biserial point discrimination? High performers will miss some questions.

Correct, high performers will miss some questions. The theory is that the high performers will miss very difficult items, not easy or moderately hard items. However, if a high performer misses a relatively easy item, it “throws off” or alters the RPBI some. The amount of alteration will be based on the total number of students. If there are 40 students in the class it might not be a big enough deal to cause a negative discrimination; however, if there are only 12 students and one of the students in the top-performing group misses an easy item (which certainly can happen) then it would cause a greater alteration in the discrimination index.

General Questions

20. How do each of you develop examinations within your programs?

[Collective practice by facilitators] We typically develop items ourselves with faculty that do a good job with item development. Items are always reviewed and edited by other faculty prior to it being on an exam. We occasionally will sit down with several instructors and go over items together to learn how to do better development. We occasionally will adapt item bank items to our exams, but those typically need some big edits. Our medical director reviews items with us prior to putting them on an exam. We have someone that blue prints the exams to make sure they are consistent with content *and* with difficulty level. We generally have two exams that we switch back and forth so that no class has the same exam as the previous class. We change out any problem item and change out approximately 10 – 15% per exam.

21. How can someone get involved with the NREMT? How do you get to go to NREMT to help and learn to write questions?

Send an email expressing interest in item writing to Suzanne Graham (sgrahan@nremt.org). She will send an application to complete and return.

22. While we utilize a professional tracking and testing program, is this considered to meet the minimum standard?

The program must still do item analysis on each major exam and evaluate how to change the exam based on the item analysis in order to meet the standard. Programs are welcome to utilize a variety of processes and products to complete item analysis procedures on high stakes exams. It is important that if you are using commercially available products that the analysis be specific to your program and your cohort of students. This information is powerful in that it informs the program on the quality of instruction, your student performance and the reliability/validity of the examination. Simply including your program's students with students from several other programs does not provide adequate program specific information.

23. For questions like knowing the dose for epi in a cardiac arrest, if that type of question is "asked" in skills performance exams do you recommend NOT including that type of question on a written exam?

That type of item should be included on the written exam, if only a small representative sample. If the program feels comfortable that the content is tested and the student is competent, that's what matters.

Programs are required to demonstrate entry-level competence for paramedic students in all three learning domains. It is important to assess students' knowledge (cognitive) ability. Furthermore, it is important to assess students' ability to utilize and apply this knowledge in context, typically done in a scenario laboratory setting, and ultimately in actual patient care encounters. Possessing the knowledge in one domain does not mean it will automatically transfer to the other, and paramedic students need to be proficient in demonstrating competence in all domains of learning.

24. What are the Angoff method and Nedelsky method?

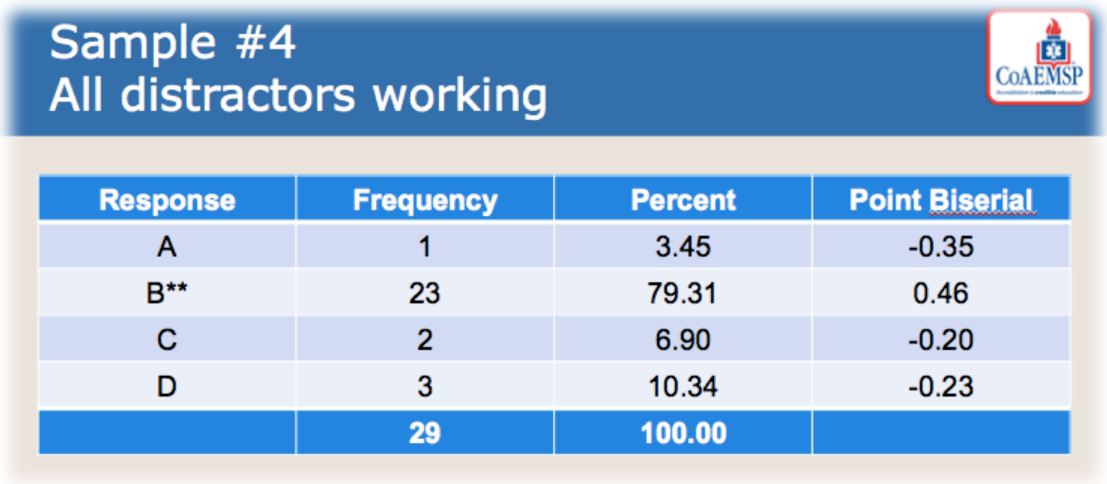
The textbook, *Foundations of Education: an EMS Approach*, 2nd ed., page 320, explains the Angoff method:

> *"An instructor can use a number of methods to predict item difficulty. The most common is the Angoff method. This method is commonly used for highstakes examinations in educational, certification, and licensure settings. The procedure is to first establish a panel of experts. The panel considers the concept of the "minimally competent candidate," or, in other words, the minimum acceptable level of knowledge. This is not the ideal or average candidate, but the candidate who is barely acceptable. The experts are then asked to estimate the percentage of minimally competent candidates who would answer that item correctly. This is done first with practice items, where the experts' estimates can be compared with actual performance of the item. As the experts rate the items, the consensus that is reached by the experts' estimates forms the Angoff rating. By computing the mean of the Angoff ratings for all items to be included in the exam, the instructor can determine a cut score. Conversion of this predicted cut score into a passing score is a matter of professional judgment for the instructor."*

The Educational Resources Information Center, which falls within the U.S. Department of Education, states this about the Nedelsky method (http://eric.ed.gov/?id=ED218361).

> *"Leo Nedelsky developed a method for determining absolute grading standards for multiple choice tests. His method required a group of judges to examine each test question and eliminate those responses, which the lowest D- student should be able to reject as incorrect. The correct answer probabilities remaining were used in computing an expected test score for the hypothetical test taker. The passing score was chosen to give the "F-D student" some probability of passing the test. Nedelsky's method of choosing passing scores are being used by many licensing and certification boards. Alternate forms of a test, containing 25 items in common, were used to measure the reliability of passing scores chosen by this method. Experts judged the tests. There were large and consistent differences between the judges for their passing scores. The experts tended to set higher passing scores the second time they judged the questions and to leave different wrong answer choices unmarked. The results of this study indicated users of Nedelsky's method should consider a two-stage judgment procedure."*

25. Sample #4. It appears we should be looking for a positive Point Biserial on the correct answer and the incorrect answers will have negative Point Biserials. Is this correct?



Sample #4
All distractors working

Response	Frequency	Percent	Point Biserial
A	1	3.45	-0.35
B**	23	79.31	0.46
C	2	6.90	-0.20
D	3	10.34	-0.23
	29	100.00	

Correct. The important number is the point biserial associated with the correct response—it's a good point biserial!