PARAMEDIC UNIT EXAMS VERSION 2

(PUE v2) TECHNICAL MANUAL

FISDAP®

USER GUIDE

Thank you for using Fisdap unit exams! These exams are great tools for assessing your students, diagnosing knowledge gaps, and identifying a plan of action for remediation. These can be used in conjunction with PUE v1, as additional practice, or new content for remediation and re-assessment. This document will give you more information about the exams, their cut scores, and how to use them. One notable change from the first version of Paramedic Unit Exams is the inclusion of a 7th unit exam – Pediatrics, which has been separated from the OB/Gyn content.

EXAM COMPOSITION

This report documents the reliability and validity of the seven Paramedic Unit Exams. The seven unit exams are: Airway, Cardiac, Medical, Ob-Gyn, Operations, Pediatrics, and Trauma. There are 75 multiple choice items on each of these exams, but the amount of items pilot tested ranged from 75-89 items per exam. The exams were administered to 254 candidates in the pilot testing phase.

Our unit exams are designed to assess knowledge, application, and critical thinking ability as students progress through their Paramedic coursework. The below table shows the composition of each unit exam.

Paramedic Unit Exam	% Knowledge	% Application	% Critical Thinking
Airway	31	41	28
Cardiology	42	40	18
EMS Ops	57	28	15
Medical	42	35	23
OB/Gyn	29	51	20
Pediatric	36	43	21
Trauma	32	45	23

EXECUTIVE SUMMARY

Validity

The Paramedic Unit Exams version 2 (PUE v2) were written and reviewed by Subject Matter Experts from all around the country. The exams require paramedic students to develop diagnoses based on a comprehensive knowledge of pathophysiology, anatomy & physiology, medical terminology, lifespan development, public health, and pharmacology.

Positive Predictive Value

This report documents the reliability and validity of the PUE v2. Each Paramedic Unit Exam is composed of 75 multiple-choice items that assess one of seven domains. The exams were pilot tested with a group of 254 paramedic students. The mean scores ranged from 52.73 to 64.94 points out of exams of that ranged from 75-89 points.

CLASSICAL ITEM ANALYSIS

The two most familiar kinds of statistics were calculated for the PUE v2 items: item difficulty and item discrimination. The average item difficulty ranges from 0.60 to .74 and the discrimination value ranges from 0.14 to 0.21. During the analysis, the difficulty and discrimination values were generally in the acceptable range. Items out of this range were flagged for further review and edited or removed from the exam.

RELIABILITY

Evidence of reliability was provided by calculating the coefficient alpha and the standard error of measurement. The overall internal consistency index as measured by coefficient alpha ranged from 0.67 to 0.81.

VALIDITY

Evidence of validity related to test content was established through the item development and review process. All items have a strong link to a certain knowledge or skill required to practice as a paramedic. Validity of score inferences is bolstered when test scores are consistent. The reliabilities of test scores were excellent, with many being in the mid to high 0.70s across subgroups. The DIF analysis with respect to gender and ethnicity helps address construct-irrelevant variance, which represents an important threat to the validity of inferences made from test scores. Results from the item-fit Rasch analysis were essentially unidimensional, providing further evidence to support interpretations based on the total test score.

Cut Score

Fisdap uses a modified Angoff method to set cut scores for each exam. In the Angoff method, a minimum of three subject matter experts (SMEs) evaluate each test item and establish the probability that a minimally competent examinee should answer the test item correctly.

These expectations were used to calculate the following exam cut scores:

Paramedic Unit Exam	% Angoff Cut Score
Airway	74
Cardiology	72
EMS Ops	71
Medical	69
OB/GYN	72
Pediatric	70
Trauma	73

OVERVIEW

Paramedic Unit Exams v2

The Paramedic Unit Exams are computer-based tests designed to prepare students for the challenges of preparing for a state or national certification examination. The selection of items for the examination were guided by standards which match those of the 2020 National EMS Practice Analysis published by the National Registry of Emergency Medical Technicians, National EMS Education Standards, and American Heart Association. More items than needed were pilot tested. The table below shows a breakdown of the exam score statistics.

Statistics	Airway	Cardiac	Medical	OB/Gyn	Operations	Pediatric	Trauma
Number of Items	87	88	86	75	86	75	89
Min	41	18	11	37	30	9	23
Max	78	79	75	68	72	63	78
Mean	64	63	59	54	61	53	65
Standard Deviation	7.56	8.37	7.94	5.88	6.19	6.89	7.19

EMS educators participated in all phases of the test development process: item writing, item selection, bias review, and data review. Fisdap organizes an advisory committee to ensure that the tests are continually informed and guided by the recommendations of content experts, measurement specialists, and EMS practitioners. The following criteria are applied to evaluate the exams: Difficulty, Precision, and Fairness.

Before they were released, the exams were pilot tested with a group of geographically diverse paramedic students. We used item response theory to measure whether the exam was a reliable assessment. In other words, we made sure that the exam was accurately able to distinguish proficient learners from weak learners.

DIFFICULTY

Prior to making the test available, the exam was pilot tested on a large sample of students. Test questions are required to achieve a statistical profile during pilot testing in order to be included on the final version of the test. The profile requires an item to be within a particular difficulty range. Items that are too easy or too difficult and therefore, provide little to no information on a student's preparedness, are omitted from the exam unless they cover essential content dictated by the test blueprint.

PRECISION

Point-biserial correlations evaluate the extent to which an item distinguishes between less proficient and more proficient learners. Items with low point-biserial scores were omitted, again with the exception of items that covered essential content.

FAIRNESS

Fisdap is committed to developing tests that are of the highest quality and are free of bias as much as possible. Fisdap tests are evaluated during development so that they do not reinforce stereotypical views of any group; are free of racial, ethnic, gender, socioeconomic, or other forms of bias; and are free of content believed to be inappropriate or derogatory toward any group.

COMPUTER-BASED TEST FORMAT

The Paramedic Unit Exams v2 (PUE2) contain 75 multiple-choice questions and cover the following seven content areas: Airway, Cardiology, EMS Operations (Ops), Medical, Obstetrics, Pediatrics, and Trauma. The items were developed by subject matter experts. This report provides evidence on the reliability and validity of these exams.

BLUEPRINT

A paramedic's responsibilities range from primary scene survey to transfer of care at a receiving facility while providing advanced life support. The included topics cover various aspects of EMS care: Airway, Cardiology, EMS Operations, Medical Emergencies, Obstetrics/Gynecology and Trauma. The demographics of the patients described in test items are balanced within each topic to reflect the appropriate percentages that are encountered in the field. 85% percent of items focus on adult patients, while 15% of items focus on pediatrics, with the exception of EMS Operations, whose items are generalized and not specific to age ranges. A paramedic is an essential link in the healthcare system and is expected to be socially harmonious while working independently on specific tasks as well as when acting as part of a larger team. This exam is designed to test the knowledge, skills and ability necessary for patient encounters, team member interactions, and the cognitive demand required for an entry level paramedic.

Classical Item Analysis

This section provides an overview of the two most familiar item-level statistics obtained from any classical (traditional) item analysis: item difficulty and item discrimination.

Statistics	Airway	Cardiac	Medical	OB/Gyn	Operations	Pediatric	Trauma
Item Difficulty Min	0.16	0.13	0.01	0.06	0.08	0.08	0.10
Item Difficulty Mean	0.74	0.72	0.69	0.72	0.71	0.70	0.73
Item Difficulty Max	1.00	0.99	0.98	0.98	0.99	1.00	1.00
Item Discrimination Min	-0.23	-0.25	-0.06	-0.06	-0.14	-0.15	-0.12
Item Discrimination Mean	0.18	0.21	0.18	0.14	0.14	0.21	0.16
Item Discrimination Max	0.52	0.47	0.37	0.42	0.40	0.66	0.43

ITEM DIFFICULTY

Prior to making the PUE2 available, each test item is required to fit within a range of difficulty. Items that are too easy or too difficult and therefore, provide little to no information, were omitted unless the item covered essential content dictated by Subject Matter Experts.

Item difficulty is noted as the average percentage of time that the question was answered correctly. Item scores are summed and then divided by the total number of students. This is also known as the p-value and ranges from 0.0-1.0. For example, if an item has a p-value of 0.92, it means that 92% of the students answered the item correctly. An item difficulty of this value suggests that: 1) the item was relatively easy, and/or 2) the students who attempted the item were relatively high performers.

RELIABILITY

Reliability

Alpha indicates the internal consistency over the responses to a set of items by measuring an underlying trait. Alpha can be thought of as the correlation or consistency between scores if the students could be tested twice with the same instrument, without the second testing being affected by the first. Alpha varies from 0 to 1.0. Higher Alpha values are more desirable for a high-quality exam.

COEFFICIENT ALPHA

A frequently reported reliability index is a Coefficient Alpha (Cronbach, 1951). The reliability coefficient is a unit-free indicator that reflects the degree to which exam scores are free of measurement error. A value of 0.70 or higher indicates an acceptable level of reliability, while values above 0.80 are good, and values above 0.90 are excellent.

VALIDITY

Validity

The Standards for Educational and Psychological Testing provide a framework for describing the sources of evidence that should be considered when evaluating validity. These sources include evidence based on: 1) test content, 2) the internal structure of the test, and 3) the relationships between test scores and other variables. In addition, when IRT (item response theory) models are used to analyze assessment data, validity considerations related to those processes should also be explored.

CONTENT VALIDITY

Test content validity evidence is provided throughout the test development process. The assessments are intended to measure students' knowledge and skills as reflected in the corresponding job competencies. A national group of subject matter experts participated in all phases of the test development and established the link between test items and what students should do in each situation as required by the profession. Each individual item written by a subject matter expert was designed to measure a specific skill. After the items were developed, they went through multiple rounds of reviews. Item alignment was also reviewed and validated by different subject matter experts and a practicing Emergency Medicine physician.

The efforts made to ensure content validity are summarized below:

- Established detailed test and item development specifications
- Trained item writers developed high-quality items
- Aligned items with the competencies and skills required in the profession
- Ensured the items were sufficient in number and distributed across content
- Reviewed levels of cognitive complexity, cut score, and importance

CONSTRUCT VALIDITY

Item-test correlations are provided in the section on Reliability. All values were positive and of acceptable magnitude. Items with slightly lower item-test correlations were identified for further review.

Rasch Model

The item response theory (IRT) model was used for the PUE2 and has a long-standing presence in applied testing programs. The use of IRT models is a standard procedure for analyzing item response data within large-scale assessments.

DESCRIPTION OF THE RASCH MODEL

The Rasch model (Rasch, 1960) was used to calibrate PUE2 items because only multiple-choice items (MCI) were part of the assessment. The Rasch model predicts the probability of person answering items correctly. It places both student ability and item difficulty (estimated in terms of log-odds or logits) on the same continuum.

UNIDIMENSIONALITY

Dimensionality is an aspect of construct validity that must be investigated when Rasch is used. This is required because IRT models assume that a test measures only one latent trait (unidimensionality). Rasch models assume that one dominant dimension determines the difference among students' performance.

RASCH ITEM FIT

Item-fit statistics (infit and outfit) were used to evaluate the degree to which the Rasch model predicts the observed item responses. Each item-fit statistic can be expressed as a mean square (MnSq) statistic. Both infit and outfit MnSq are the average of standardized residual variance (the difference between the observed score and the Rasch estimated score divided by the square root of the Rasch model variance). The infit statistic is weighted by the examinee locations relative to item difficulty and tends to be affected more by unexpected responses close to the person, item, or rating scale category measure.

Reasonable values for MnSq should range from 0.5 to 1.5 and for high-stakes testing. MnSq values close to 1.0 are desirable. Deviation in excess of the expected value can be interpreted as 'noise' or lack of fit between the items and the model.

Differential Item Functioning

Items that lack bias provide further evidence that inferences made from assessment results are valid. For this purpose, the PUE2 items were analyzed for differential item functioning (DIF). Empirical investigation of DIF strengthens the validity evidence related to score interpretations for students in particular groups by eliminating potential sources of construct-irrelevant variance. DIF occurs when examinees with the same ability level but different group memberships do not have the same probability of answering the item correctly. This pattern of results may suggest the presence of item bias. As a statistical concept, however, DIF can be differentiated from item bias, which is a content issue that can arise when an item presents negative group stereotypes, uses language that is more familiar to one subpopulation than to another, or is presented in a format that poses a disadvantage to certain learning styles. While the source of item bias is often plain to trained judges, DIF may have no clear cause. However, studying the way that DIF arises and how it presents itself has an effect on how to detect and correct it.

LIMITATIONS OF STATISTICAL DETECTION

No statistical procedure should be used as a substitute for rigorous, hands-on reviews by content and bias specialists. The statistical results can help organize the review so that the effort is concentrated on the most problematic cases. Furthermore, no items should be automatically rejected simply because a statistical method flagged them; or accepted because they were not flagged.

Statistical detection of DIF is an inexact science. There have been a variety of methods proposed for detecting DIF, but no single statistic can be considered either necessary or sufficient. Different methods are more or less successful depending on the situation. No analysis can guarantee that a test is free of bias, but almost any thoughtful analysis will uncover the most flagrant problems.

A fundamental shortcoming of all statistical methods used in DIF evaluation is that all are intrinsic to the test being evaluated. If a test is unbiased overall, but contains one or two DIF items, any method will locate the problems. If, however, all items on the test demonstrate consistent DIF to the disadvantage of a given subpopulation, a statistical analysis of the items will not be able to separate DIF effects from true differences in achievement.

Universal Design and Test Accommodations

The goal of universal design in test development is to maximize accessibility without adaptation or special design. The application of universal design principles offers a test that increases the participation of all students, including those with disabilities and English Language Learners. In practice, universal design considers the needs of different subpopulations to maintain test fairness.

Fisdap employs the following universal design principles: Principle Guidelines

Equitable Use: Provide the same means of use for all users. Avoid segregating or stigmatizing users. Provide equal availability for privacy, security, and safety. Make the design appealing to all.

Flexibility in Use: Facilitate the user's accuracy and precision. Provide adaptability to user's pace.

Simple and Intuitive: Eliminate unnecessary complexity. Be consistent with user expectations and intuition. Accommodate a range of literacy and language skills. Arrange information in order of importance. Provide effective prompting and feedback.

Perceptible Information: Use pictorial modes for presentation of essential information. Provide adequate contrast between essential information and its surroundings. Differentiate elements in ways that can be easily described.

Tolerance for Effort: Arrange elements to minimize hazards and errors. Provide warnings and fail-safe features. Discourage unconscious action in tasks that require vigilance.

Fisdap incorporated these design principles into item development for the PUE2. The standardized Fisdap test applies universal design practice when reviewing item content and test booklet layout against these design guidelines.

Fisdap's universal design guidelines were implemented in item development for the PUE2. The following considerations are incorporated in the Fisdap item development training materials for contributing SMEs.

I. Tests

1. Fairly represent as many groups as is reasonable.
2. Include many perspectives characterized by an issue rather than presenting only one side.
3. Balance the roles for the groups represented. For example, include the contributions of both males and females as well as of various ethnic minority groups.

II. Considerations for items – What to Avoid:

1. Descriptions of groups in terms of physical, personality, or interest stereotypes
2. The use of language that might be considered derogatory to any group
3. The use of words that have different meanings in different cultural settings or dialects
4. The use of subject matter likely to be unfamiliar to some groups, while familiar to the majority
5. The use of esoteric vocabulary or complex sentence structure when that is not what is being tesed

III. Considerations for items – What to Do:

1. Include material relevant to and stressing the positive aspects and values of diversity
2. Present positive role models from various groups or material that discusses the contributions of groups to science, history, government, and the arts

Concepts of universal design are also incorporated in the graphic design of the PUE2 online tests, which include:

PRODUCTION

1. Use a font style that is easy to read
2. Enlarge the font size
3. Allow large space between items, frame items for easy identification, and use graphic item labels

ADMINISTRATION

1. Provide adequate testing time
2. Repeat instructions

Each PUE2 item was designed based suggested universal design item features.

VI. Content concerns

1. Every item should reflect specific content and a single specific mental behavior, as called for in test specifications (two-way grid, test blueprint).
2. Base each item on important content to learn; avoid trivial content.
3. Use novel material to test higher level learning. Paraphrase textbook language or language used during instruction when used in a test item to avoid testing for simply recall.
4. Keep the content of each item independent from content of other items on the test.
5. Avoid over specific and over general content when writing MC (multiple choice) items.
6. Avoid opinion-based items.
7. Avoid trick items.
8. Keep vocabulary simple for the group of students being tested.

VII. Formatting concerns

1. Use the question, completion, and best answer versions of the conventional MC, the alternate choice, true-false (TF), multiple true-false (MTF), matching, and the context dependent item and item set formats but AVOID the complex MC (Type K) format.
2. Format the item vertically instead of horizontally.

VIII. Style concerns

1. Edit and proof items.
2. Use correct grammar, punctuation, capitalization, and spelling.
3. Minimize the amount of reading in each item.

IX. Writing the stem

1. Ensure that the directions in the stem are very clear.
2. Include the central idea in the stem instead of the choices.
3. Avoid window dressing (excessive verbiage).
4. Word the stem positively, avoid negatives such as NOT or EXCEPT. If negative words are used, use the word cautiously and always ensure that the word appears capitalized and boldface.

X. Writing the choices

1. Develop as many effective choices as you can, but research suggests three is adequate.
2. Make sure that only one of these choices is the right answer.
3. Vary the location of the right answer according to the number of choices.
4. Place choices in logical or numerical order.
5. Keep choices independent; choices should not be overlapping.
6. Keep choices homogeneous in content and grammatical structure.
7. Keep the length of choices about equal.
8. Avoid **All-of-the-above** or **None-of-the-above**
9. Phrase choices positively; avoid negatives such as NOT.
10. Avoid giving clues to the right answer, such as
    i. Specific determiners including **always, never, completely,** and **absolutely**.
    ii. Clang associations, choices identical to or resembling words in the stem.
    iii. Grammatical inconsistencies that cue the test-taker to the correct choice.
    iv. Conspicuous correct choice.
    v. Pairs or triplets of options that clue the test-taker to the correct choice.
    vi. Blatantly absurd, ridiculous options.
11. Make all distractors plausible.
12. Use typical errors of students to write your distractors.

Software and Estimation Algorithm

Analyses were implemented via jMetrik, R (package Psych). Rasch item calibrations were centered on the items with a mean = 0 and scale = 1.

References

AERA, APA, NCME (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education). (1999). *Standards for Educational and Psychological Tests*. Washington, DC: American Educational Research Association.

Cronbach, L. (1951). Coefficient Alpha and the Internal Structure of Tests. *Psychometrika*, 16, 297–334. Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research. Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22, 719–748.

Meyer, J. P. (2014). Applied Measurement with jMetrik. New York: Routledge.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Copenhagen: Danish Institute for Educational Research.