Marching Band Pedagogical Practices: An Assessment and Statistical Analysis

Zachary W. Eisengrein

Department of Music Education, University of Miami

MED 764: Music Assessment

Dr. Stephen Zdzinski

November 27, 2023

Table of Contents

Marching Band Pedagogical Practices: An Assessment and Statistical Analysis

This project contains a full assessment of student knowledge and application of marching band practices and pedagogies as well as the data for the processes of creating, administering, evaluating, and reflecting upon the test. The purpose of this project is twofold: firstly, it is designed to assess student knowledge of the stated topic. Included in that topic is field terminology, marching fundamentals, instrument maintenance and repair, general ensemble practices, and practices specific to marching band instruction. These can all be seen in Table 1, the Table of Specifications, as well as the point values and taxonomy levels assigned to each area of evaluation. The second purpose of this project is to allow me the opportunity to engage with the lengthy process of assessing knowledge of a particular construct, from test conception to test evaluation and everything in between. This essay serves as a report for all of the creation processes, statistical findings, implications, and reflections upon future revisions.

Table of Specifications

Below is the table of specifications outlining the content areas measured in this assessment as well as their corresponding taxonomy levels and point values. The test, worth up to 70 points, is broken into point values by taxonomy: 24 points for "knowledge," 6 points for "understanding," 27 points for "application," 3 points for "analysis," and 8 points for "evaluation" of the content areas. Two points are granted if the test-taker provides their name and the date on the test.

Table 1

Specifications for Content and Taxonomy Levels Measured

Content Measured	**Taxonomy Level – Knowledge** (24 points)	**Taxonomy Level – Understand** (6 points)	**Taxonomy Level – Apply** (27 points)	**Taxonomy Level – Analyze** (3 points)	**Taxonomy Level – Evaluate** (8 points)
Field Terminology (9 questions)	8 matching 1 T/F (5pts)	N/A	N/A	N/A	N/A
Marching Fundamentals (5 questions)	1 matching, 1 T/F (8pts)	N/A	3 short answers (9pts)	N/A	N/A
Instrument Maintenance (4 questions)	2 multiple choice (2pts)	N/A	2 multiple choice (6pts)	N/A	N/A
Ensemble Practices (11 questions)	3 multiple choice, 2 T/F (5pts)	2 short answers (4pts)	2 short answers, 1 multiple choice (9pts)	N/A	1 essay (6pts)
Marching Band Pedagogy (6 questions)	3 T/F (3pts)	1 multiple choice (2pts)	1 multiple choice (3pts)	1 short answer (3pts)	N/A

70 total pts = 24 + 6 + 27 + 3 + 8 + 1 + 1

Student Sample Description

Considering the purpose of the assessment, to assess knowledge and application of skills within marching band pedagogy, I had to first consider a population that would be most appropriate for receiving it. In order to get the most accurate representation of student knowledge with the most discriminating set of scores, I decided the population was to be a mix of participants whose levels of involvement in marching band were varying. The sample comprised of 15 college-level students, both undergraduate and Graduate, with varying levels of experience and involvement in marching band. Of the 15, 14 were involved in marching band while 1 was not but was a student in music education. 5 of the 14 involved in marching band were members of marching band leadership at the time of study, and 3 were marching band Graduate Teaching Assistants. Slightly more than half (8 out of 15) of all test-takers desired to go into the field of music education, while 7 out of 15 desired to become teachers of marching band specifically.

Test Construction Procedures

In planning this assessment, perhaps the most difficult aspect of the beginning stages was deciding what to assess. In a classroom setting that has a structured curriculum, there is much less mystery when it comes to knowing what is to be assessed and when. Without the classroom structure, however, I decided it would be best if I chose a topic I cared about, such as teaching marching band. In order to create a marching band assessment that would be both an accurate representation of student knowledge as well as highly discriminating, I chose to first assess the areas of marching band pedagogy that are most common and relevant to students learning to become band directors. The content areas, which can be seen in Table 1, include knowledge of field terminology, marching fundamentals including step size and marching techniques, principles of basic instrument maintenance, general ensemble practices,

as well as marching band-specific pedagogical practices. While these content areas are intended to reflect particular aspects of students' knowledge, understanding, application of, analysis of, and evaluation of the content, they are also intended to help create a wholistic measure of students' overall proficiency in marching band pedagogy.

The final test product was created after several significant revisions. In the early stages, I struggled most to create questions that all measured one underlying construct, and rather created questions where there seemed to be no single distinction as to what the objective was; the first drafts appeared to be more a measure of a general knowledge of random marching band topics than specific principles of pedagogy. While I gave myself grace, considering this was my first attempt at creating test questions, I quickly realized that the test needed to have a focal point rather than just be a rapid-fire, jeopardy-style marching band knowledge test. After many revisions, the intent to wholistically measure proficiency in the subject area by breaking it down into individual categories became much clearer. Furthermore, in order to exercise multiple levels of learning, I incorporated all item types into the assessment including matching, short answer, multiple choice, true or false, and an essay with the goal of assessing the "knowledge," "understanding," "application," "analysis," and "evaluation" taxonomies.

Descriptive Statistical Summary

Table 2

Raw Test Data and Calculations of Central Tendency and Dispersion

Raw Scores (%, highest to lowest)	100 95.7 95.7 95.7 95.7 94 93 93 93 91 89 86 78.5 77 65
Mean	89.49
Median	93
Mode	95.7
Range (points)	35
Variance (points^2)	86.66
Standard Deviation	8.99 points from the mean

After administering the test, the most immediate statistical findings were the raw scores seen in the top row of Table 2. The raw scores are unaltered scores that represent only how many points students achieved out of the total number of points possible. In this case, they are percentages that are proportionate with the number of points achieved. Most students achieved raw scores that were high; the median was 93%, the mode was 95.7%, and the mean was 89.49%. With a mean score that is lower than both the median and the mode, a negative skew is indicated meaning that most scores fell on the upper side of an asymmetrical distribution. In addition to the calculations of central tendency, calculations representing the dispersion of scores were also made: the scores, on average, deviated 8.99 points from the mean score of 89.49, with a variance of 86.66. Because the variance is so close to the mean, a relatively low spread of scores is indicated. If plotted on a graph, with the variance, deviation, and skew in mind, the majority of scores would fall to the right of the curve (the mean) and be clustered closely together, all of which is supported by the data in Table 2 and the calculations laid out in Appendix B.

Item Analysis

In addition to determining central tendencies and dispersion of test scores, an item analysis consisting of calculations for item difficulty and item discrimination was conducted. Together, these two components of analysis provide a powerful, quantified reflection on the test items created as they suggest those which are effective and those in need of revision. In order to determine item difficulty, the formula P = R/T was used for all items, where R is the number of correct responses from the sample group and T is the total number of responses. In this case, T was always equal to 15. Findings concluded that 50% of all test items received a difficulty rating of 1.0, meaning that all students who answered these questions answered them correctly. The other 50% of items received difficulty ratings ranging from 0.933 to 0.6, and the average item difficulty for those was 0.86. All calculations and detailed findings for item difficulty can be found in Appendices B and C.

The item discrimination index (DI) is used to understand the proportion of students in the highest scoring group who answered each item correctly to the students in the lowest scoring group who answered each item correctly. In order to determine the highest and lowest scoring groups of students, I separated the total population of test-takers into three categories, each containing 5 students: there were 15 total test takers, so the highest 33% were those who scored 95.7% and above, and the lowest 33% were those who scored 89% and below. The middle 33%, or those who scored from 91% to 94%, were not used to calculate discrimination.

Using the highest and lowest groups of scorers, the DI was calculated by dividing the number of students in the highest 33% who got each item correct by the number of students in the lowest 33% who got each item correct. The closer the DI is to 0, the less discriminating the item. After completing the item analysis for discrimination, it was found that, similar to item difficulty, 50% of all test items were non-discriminating, receiving score of 0. This

means that all students in both the highest and lowest 33% of test takers answered these items correctly. For the remaining 50% of items, discrimination values varied from 0.2 to 0.6, with the average DI being 0.3. Only 7 of the 21 items in the discriminating 50% of test items were found to be *highly* discriminating, receiving a DI of 0.3 or higher. The greatest DI was 0.6, occurring in only three items. All calculations and detailed findings for item discrimination can be found in Appendices B and D.

For the sake of gaining a deeper understanding of the test items I created, I made further calculations related to reliability which were intended to provide a more wholistic image of item effectiveness by determining how consistently the items are assessing the underlying construct. I chose to calculate reliability using split-half reliability and the Pearson Correlation formula. This was the most appropriate choice over test-retest reliability and inner-rater reliability because the test was only administered once by one "researcher" (myself). Split-half reliability is an assessment of internal consistency done by dividing the items into two even halves – by number, by difficulty, even and odd, item type, etc. – and calculating how effectively each half is measuring the construct in relation to the other.

Ideally, the halves of the test would be equally reliable and consistent in their assessment, meaning there is a strong correlation between the first half's assessment and the second half's assessment of the material. To find the split-half reliability, I first divided my test in half by number of items. The first group, defined as the *x*-variables, are items 1 through 21 (numbers 1-14), and the second group, defined as the *y*-variables, are items 22-42 (numbers 15-35). On the test, items are not numbered 1-42; question 9 is a matching question that involves 8 parts and is listed as 9a-9h, which can be referenced in Appendix A. After diving the items, I then averaged students' *new* raw scores (their scores on just the *x*- and *y*-halves of the test) to calculate correlation using the Pearson Correlation formula. The Pearson Correlation formula was chosen over the Spearman-Brown formula for reliability because it

measures the strength of a linear relationship between two sets of data, an *x* and a *y*. The calculations yielded a correlation coefficient of -0.10. For the greatest correlation, the coefficient should be as close to ± 1 as possible, which would be near-perfect. Anything within ± 0.50 of ± 1 is considered to be a strong correlation, thus meaning that -0.10 is an extremely weak correlation. A weak, negative correlation such as this indicates that when one set of variables increases, the other decreases drastically. All data regarding the calculation of internal consistency using the Pearson Correlation formula can be seen in Appendix E.

Discussion

Creating, administering, and evaluating a test has proven to be a much lengthier and more involved process than it is given credit for. In the creation of this project, I encountered many surprises that furthered my position as a test-creating novice, and many of these surprises came when encountering the realities, whether in truth or falsehood, of my preconceived notions and biases regarding the assessment procedures. While, conceptually, I understood that creating an effective test would be challenging, I endured significant struggles when building a set of items that would not only appropriately assess what they were intended to, but also fit within one specific "theme" or purpose. As a result, my initial drafts consisted primarily of items testing rote memorization of random marching band facts and principles rather than items that were connected by a common assessment goal. Such ambiguity in the first several drafts forced me to reconsider, what was I assessing *for?* What was the intention and objective behind creating this assessment? That eventually led me to the more solidified, wholistic concept of marching band pedagogical practices, including knowledge of field terminology, marching fundamentals, instrument maintenance and repair, general musical ensemble practices, and teaching topics specific to leading a marching band. In addition to the goal of the test itself, this entire project has also served as an opportunity for me to better learn how to create, administer, and evaluate an assessment. My experience with the process and all its surprises is detailed in this section.

The first surprise I encountered, other than the discovery of how difficult it is to create a pool of effective questions, occurred when the test was administered to my student sample. While students were taking the test, several of them brought to my attention the subjective nature of some of the items. These were questions that either contained terms not common or familiar to all marching band programs or that tested philosophical concepts rather than known facts. Each of these issues are surely a result of not only my own bias, but also my

inexperience as a teacher, as it had not occurred to me that some programs did not use the same terms that I was familiar with. Such terms include “side 1” versus “side A” and “side 2” versus “side B” in questions 5 and 7, respectively, “flanks” and “obliques” in question 9, and “zero points” in question 32. In addition, question 18, asking students to choose the most appropriate order of balance in ensemble playing, was reported to be an issue of either philosophical subjectivity, contextual subjectivity, or both. Each of these are issues concerning the overall validity of my test; the unfamiliar terms and/or concepts are likely to lead to guessing, which results in an inaccurate measure of what the assessment was intended to capture. If I were to rewrite these items, I would attempt to be more specific in my instructions and expectations and use terms and concepts that are universal to marching bands rather than specific to one or two programs.

Another surprise I encountered came in the process of item analysis. In my analysis, I calculated both the difficulty rating and the discrimination index (DI) for all items. What I was expecting to discover was that most of the questions would discriminate to the desired extent or greater. However, this was not the case, and actually it was quite the opposite; while the ideal testing item would have had a medium-to-high difficulty with proportionately high discrimination, my testing items instead were a combination of high and low difficulties with high and low discriminations, providing varying levels of accuracy in their measurement. In fact, 50% of all my test items were neither difficult *nor* discriminating (they received a difficulty rating of 1.0 and a DI of 0). This was very surprising to me, as it implicated that 50% of the questions were easily answerable by students in the highest, middle, *and* lowest scoring groups, thus providing a less than accurate measurement of the extent of their true, wholistic knowledge.

For the remaining 50% of items, difficulty and discrimination varied widely and, in some cases, were a misrepresentation of true student knowledge. Such items were those with

validity concerns such as numbers 5, 7, 9, and 32, which were discussed previously. Many of the remaining questions provided low difficulty ratings with low discrimination, however there were a few outliers that were surprisingly discriminating, surprisingly difficult, neither of those, or an unusual balance of the two. Items 15 and 33 were among those that were intended to be most difficult, yet the combination of their high difficulty rating with a *low* DI suggest that they were actually *too* difficult considering that students in both the highest and lowest scoring groups got the item incorrect by majority. If this were material being taught in a classroom, the intelligent instructor would use this data to inform their future teaching or assessment-building, whether that means teaching the materials differently, creating the test differently, or being more specific in their expectations (in other words, letting students know what concepts would be tested ahead of time!). Only a handful of testing items proved to be appropriately difficult with an equally appropriate DI and did not involve issues of validity, such as numbers 12 and 35, for example.

With all of this in mind, I would like to reflect further on the testing items whose data was surprising to me, either in that they were surprisingly difficult, discriminating, or neither. The questions that I had expected to be the most challenging were 22, 23, 27, and 35, and of these, only items 23 and 35 proved to be difficult *and* highly discriminating. The others provided varying combinations of little-to-no difficulty with little-to-no discrimination. I also expected number 15 to be challenging yet discriminating, but as mentioned above, it was *too* challenging to be discriminating at all. Questions that proved to be surprisingly difficult included numbers 12, 24, and 33, as well as 9 and 32, although these were among the items whose validity is questionable. All individual testing items being referenced in this discussion can be found in Appendix A. All data regarding item difficulty ratings and discrimination indices can be found in Appendices B, C, and D.

Finally, while the overall validity of the assessment has shown to be a general concern when it comes to the test's overall effectiveness, this concern is deepened by the findings for the test's *reliability*, which are highly suspect. In the section above on item analysis, I detailed my process for calculating the reliability of the test items, where I used split-half reliability that was found using the Pearson Correlation formula. The goal when it comes to reliability is to have the two sets of items correlate, meaning that when one score goes up, the other goes up with it proportionally. What I found after calculating reliability was quite a shock; a Pearson Correlation coefficient of -0.10, indicating a very weak, negative correlation.

Now, given the questionable validity of my test, I knew before making any further calculations that my reliability would also be concerning. It certainly is, and more so than I could have expected, but with that said, I think it is worth noting the difference in item difficulty and taxonomy levels assessed between the two sets of data created by using the split-half reliability. The *x*-values (numbers 1-14) involved questions that were much easier and assessed lower taxonomy levels than those in the *y*-values. For example, numbers 1-9h are all matching questions at the lowest taxonomy level, "knowledge." The *y*-values, however (numbers 15-35), involved short answers, multiple choice, true/false, and an essay, most of which were more difficult questions that test skill at higher taxonomy levels, for example, "analysis," "application," and "evaluation." This segregation of difficulty has made itself highly apparent in the correlation findings. If I were to recalculate the split-half reliability, I would consider splitting the items differently to make for a more evenly distributed measure of difficulty. With that said, however, the inner consistency of a test should not necessarily depend on how the items are split; the test should, overall, be reliable and consistent within itself regardless of how the items are arranged. All data regarding reliability can be found in Appendix E.

What is interesting about the findings for reliability is how drastically different my findings were in my original calculations which were made in error. Initially, I computed the correlation with an uneven split of the test items; rather than splitting from numbers 1-14 and numbers 15-35, I split the test from numbers 1-21 and 22-42 (which is the true number of *items* if you consider the 8-part question in number 9). However, I did not compensate for 9a-9h in such a split. Instead, I calculated correlation with my x-values being data from numbers 1-21 and my y-values being data from only numbers 22-35. While I recognize my error and obviously have since corrected it, I think it is worth noting how drastically different the results were when the Pearson coefficient was found the first time. With this flawed data set, the correlation coefficient came out to be -0.99, indicating a very strong, though still negative, and nearly perfect correlation between the two data sets. Given the validity of my test which was found and discussed at length previously, I was highly suspicious of such optimistic findings for reliability. Having made the recalculations, the results make much more sense and only deepen my sense of understanding for how truly demanding it is to create a test that is accurate, effective, and consistent.

Recommended Revisions of the Measure

With the prior discussion in mind, the following bullet-pointed revisions are those I would consider if this assessment were to be revised and readministered:

1. Refrain from creating and using items whose correct answers are based on issues of philosophical and/or contextual subjectivity rather than known facts.
2. In numbers 1-8, be more specific in the field diagram and expectations for/use of terminology. This would aim at eliminating confusion about the terminology and orientation of the diagram.
3. In general, I believe that being more specific about expectations and instructions would have contributed to greater student success.
4. Use the most universal terminology available to deter guessing as much as possible. If no single term is proven to be the most universal, then provide multiple terms to define the same concept, such as writing "Flanks/Snap Turns/etc." instead of solely using "Flanks" in item 9, for example.
5. Revise the questions that demonstrated high difficulty but then had low discrimination. Strategies for doing so may include being more specific in stating expectations and instructions.
6. Revise the entire assessment further by creating a more discriminating pool of items. Having 50% of test items be totally non-discriminating may raise concerns as to whether the test is valid and/or reliable in its measure.
7. Define the purpose of the assessment more precisely for students so they know what they are being tested on (wholistically *and* specifically!) and why.
8. Not that it should matter how the questions are split, but consider calculating reliability with different split-half data sets to see if different results are found.

Conclusions and Implications

In creating this project as a first-time test-maker, I have learned that there is a *lot* that goes into building an assessment that is effective. I have learned that there is much more work that I will have to do in the future to ensure a more valid and reliable measure of the underlying construct I am assessing, and this includes being more specific in my instructions and expectations, eliminating any subjectivity, and working harder to create more discriminating items. In most cases, I would consider the effectiveness of this assessment to reflect the average item difficulty and discrimination index, in part, as those provide the most specific data about the effectiveness of each individual item. However, in this case, I have chosen not to calculate the average difficulty nor the average discrimination, as the representations of each would be invalid due to the issues detailed in the discussion above regarding validity and reliability. With that being said, I think this has still been a very valuable learning experience and I am eager to continue learning to be more effective in my assessments of student knowledge.

When it comes to statistical conclusions, I have attempted to examine whether there is a correlation between students' involvement in marching band and their performance on this test. Obviously, any involvement in marching band will guarantee success over those with no involvement, and this is supported (though only by one case) by the raw scores of students who *do* have experience versus the single student in the sample who had zero experience and scored a 65%. However, a more in-depth study testing multiple population groups would have to be conducted in order to determine any sort of correlation between the extent of one's marching band involvement and their potential to do well on the assessment. Before such in-depth studies could be conducted at all, however, a more valid and reliable assessment would be necessary to determine correlation and the extent of impact on test scores between student involvement and student performance on the assessment.

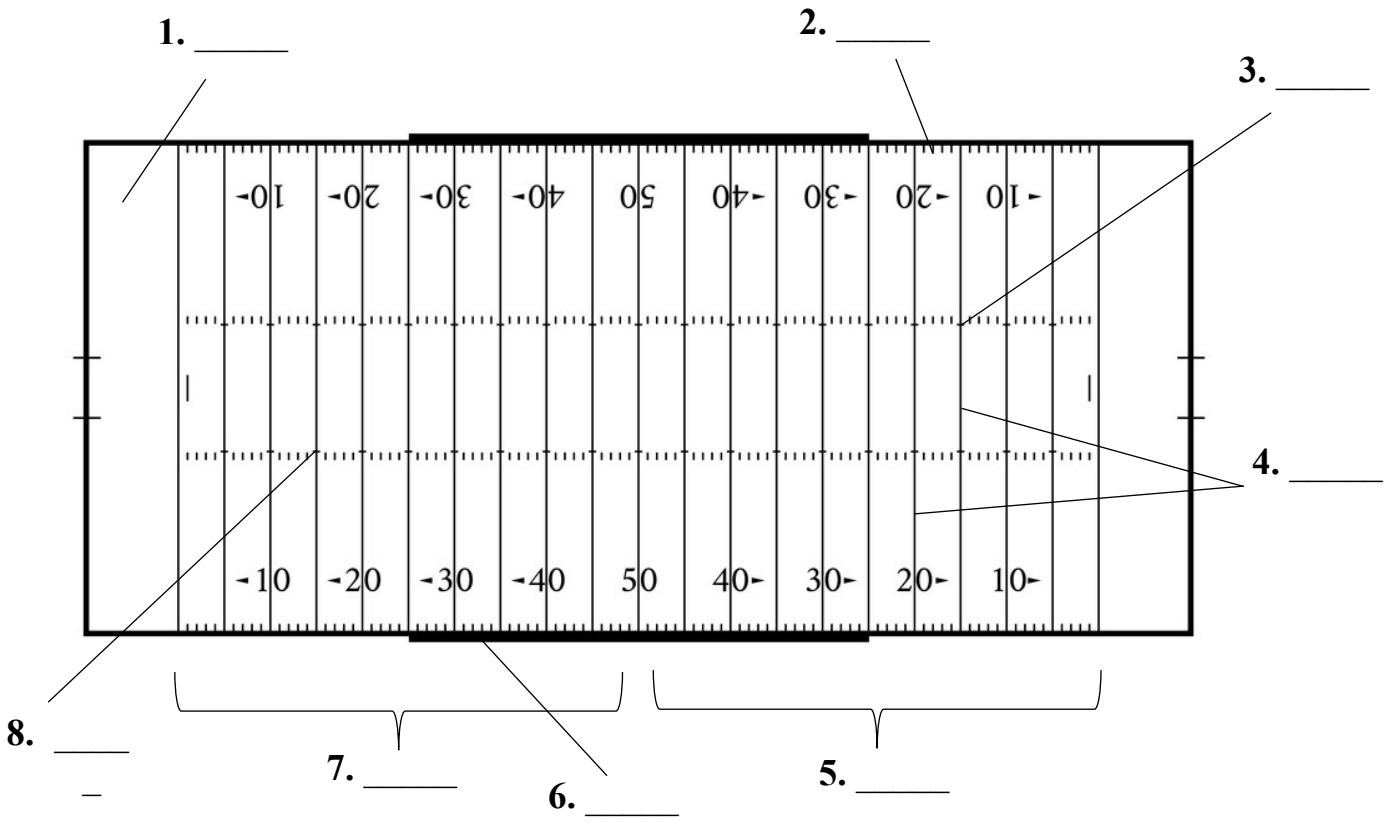
Appendix A

Sample Test Administered to Students

Font size and spacing in the sample test below has been minimized.

Part I: Matching

Directions: Identify the parts of the field by selecting the most accurate term from the word bank. **(4 points total, 0.5 pts each)**



Word bank

Yard Lines	Touchdown	Side A	Side 1	Back Hash
Front Sideline	Endzone	Band Line	Field Lines	Side 2
Front Hash	Back Sideline	Side B	Band Hash	Numbers

9. Match each marching fundamental with its description. **(8 pts total, 1 pt. each)**

_____ Forward March

_____ Backwards March

_____ Mark Time

_____ Attention

_____ Parade Rest

_____ Slides

_____ Flanks

_____ Obliques

A. Marching in a diagonal direction while the upper body stays flat to the sideline. Uses a slightly larger step size than a forward or backward march.

B. Can be called to this position using whistles or hand claps, often in response to a "ten-hut" or similar command.

C. Use of peripheral vision to preserve the integrity of a line or shape in the drill.

D. The presence between relaxed and attention. Feet are apart with horns or sticks in a rested position. Band is still focused in this position.

E. Moving linearly across the field where the step is initiated from the platform and rolled through from heel to toe.

F. Moving linearly across the field so that steps are initiated from the platform and are not rolled through.

G. Sharp transition from marching forward in one direction to marching forward in another direction.

H. Moving the feet in time without directional motion.

I. Feet march in one direction, forward or backward, while the upper body stays flat to the sideline.

Part II: Short Answer

Directions: Using the information provided, fill in the blanks with the correct answer to each of the following questions. **(22 points total)**

10. If an 8-to-5 step size indicates taking 8 steps for every 5 yards on the field, then what would a 26-to-5 step size indicate? **(3 pts)**

_____

11. If you are marching a perfect 8-to-5 step size at 120bpm, how many steps would you have to take to travel from the 50 yard-line to the endzone? **(3 pts)**

_____

12. If you march exactly 60 steps from the 50 yard-line at a 12-to-5 step size, where would you end up on the field? **(3 pts)**

_____

13. Describe one strategy for effectively dressing "diags" in a drill form. **(2 pts)**

_____
_____
_____
_____

14. Why would it be important for the band to understand who has the melody at any given point in the music? **(2 pts)**

_____
_____
_____
_____

15. Which of the following keys would be the **most impractical** for a marching band arrangement: B Major, Ab Minor, C Minor, D Major, or Eb Minor? Give at least two reasons why. **(3 pts)**

_____
_____
_____
_____

16. Look at the drill form pictured below. Identify one way it could be improved and suggest one strategy the band might implement during rehearsal to improve that issue. **(3 pts)**

17. In many college pre-game shows, it is tradition for the band to "float" the school's letter or icon across the field such as in the image pictured below. Describe at least two strategies the band can utilize in order to preserve the drill form while moving across the field. **(3 pts)**

Part III: Multiple Choice

Directions: Choose the best answer to each of the following questions regarding ensemble practices, instrument repair, and pedagogy. **(19 points total)**

18. The most accurate order of ensemble balance when playing in the marching band from most prominent to least prominent is **(1 pt.)**
   a. High voices, middle voices, low voices
   b. All voices should play at equal volumes
   c. Low voices, middle voices, high voices
   d. Low voices, high voices, drums

19. When the ensemble is playing together, who should lead a crescendo or decrescendo? **(1 pt.)**
   a. The loudest voices
   b. The lowest voices
   c. Section Leaders
   d. The Band Director

20. If trumpets are playing a G and an A at the same time, how should they approach tuning that harmony? **(3 pts)**
   a. By tuning the notes really close to each other
   b. By playing as loudly as possible
   c. By doing what their Section Leader is doing
   d. By "pulling" the notes apart from each other

21. What is the most reliable way of starting a song together as an ensemble while performing on the field? **(1 pt.)**
   a. By listening and responding when the rest of the band begins playing
   b. By doing what their section leader is doing
   c. By listening and responding to the conductor's whistles
   d. By following and responding to the conductor's hands

22. What would be the most appropriate sequence for teaching marching fundamentals? **(3 pts)**
   a. Attention/Parade Rest, Mark Time, Forward March, Slides
   b. Forward March, Backward March, Slides, Mark Time
   c. Attention/Parade Rest, Slides, Obliques, Backward March
   d. Forward March, Flanks, Slides, Attention/Parade Rest

23. When tuning a drumhead, what is the most appropriate method of tightening or loosening the nuts so as not to stretch or warp the head unevenly? **(1 pt.)**
   a. Tuning in a crisscross pattern
   b. Tuning in a clockwise pattern
   c. Tuning in a counterclockwise pattern
   d. Tuning in an unplanned order of your preference

24. If a key on a saxophone is loose and floppy, what is the most likely cause?**(3 pts)**
   a. A screw is loose somewhere on the instrument
   b. A pad is sticky somewhere on the instrument
   c. The body of the instrument is bent somewhere
   d. A spring is out of place somewhere on the instrument

25. If a trumpet's second valve does not play immediately after oiling and reinserting it, what is the most likely cause? **(3 pts)**
   a. The valve was not oiled enough
   b. The valve was reinserted incorrectly
   c. The player is not using enough air
   d. The valve is bent

26. Why might a significant bend in an instrument's bell affect the sound it produces? **(1 pt.)**
   a. It becomes a different instrument when the bell is bent
   b. It becomes unplayable when the bell is bent
   c. The vibration pattern of the instrument changes when the bell is bent
   d. A bend in the bell would not affect the instrument's sound

27. Which of the following most accurately describes the purpose of "check, adjust, run it back" when learning and rehearsing drill? **(2 pts)**
   a. It provides students the chance to self-assess and self-diagnose mistakes
   b. It provides students the chance to criticize the performance of their peers
   c. It provides students the chance to memorize the music
   d. It provides students the chance to understand how they fit into the "big picture" of the drill

Part IV: True or False?

Directions: Select whether each of the following statements is true or false by circling the corresponding initial (T = true, F = false). **(7 points total, 1 pt. each)**

28. When playing, bell-front brass will be easily heard regardless of the direction they are facing. **T / F**

29. When rehearsing drill or music, the bass line should position themselves so that the heads of their drums are facing towards the audience or Director. **T / F**

30. When holding a brass instrument at attention, the mouthpiece should be at eye level. **T / F**

31. The woodwind instrument that most often doubles the trombones in an arrangement is the alto saxophone. **T / F**

32. Zero points on the field denote four-step spacing at a 4-to-5 step size. **T / F**

33. The invention of the Sibelius software greatly innovated the ability of marching band directors to create, deliver, and teach drill to their students. **T / F**

34. The woodwind instrument that most often doubles the Bb horn or mellophone in an arrangement is the alto saxophone. **T / F**

Part V: Essay

Directions: Answer the following prompt in at least one paragraph using your own knowledge of marching band rehearsal practices.

35. *Directions:* Explain why sound delay occurs and how it can create challenges for marching ensembles. Provide an example of how this might occur during a rehearsal or performance. Additionally, offer at least two suggestions for how you, as a member of the ensemble, could combat such challenges using your knowledge of rehearsal practices and the nature of sound delay. **(8 pts)**

Essay Rubric

Content	4 – Outstanding Demonstration of Knowledge	3 – Acceptable Demonstration of Knowledge	2 – Lacking Demonstration of Knowledge	1 – Little to No Demonstration of Knowledge
Sound Delay	Student explains the concept of sound delay efficiently and accurately in the context of marching band and provides at least one clear and accurate example of how it might appear during rehearsals and/or performances.	Student describes sound delay but provides a slightly unclear example of how it might appear in marching band by excluding specific marching band-related terms or concepts.	Student's explanation of sound delay is not totally accurate or efficient. Student's example of how it may appear in the marching band context was also not totally clear or accurate.	Student gave an inaccurate explanation of sound delay, or no explanation at all, with an inaccurate example of how it might appear in marching band.
Pedagogical Approaches	Student provides at least two outstanding and accurate examples of how sound delay may be combatted during rehearsal using concepts learned such as who is in charge of time keeping and how location on the field affects exactly when one plays.	Student provides two examples of how to combat sound delay which are accurate but supported by few details regarding concepts learned in class and rehearsal.	Student provides at least one example of how to combat sound delay, but the description of specific application techniques was not totally accurate or clear. Student did not demonstrate full knowledge of the concept.	Student provided one or fewer inaccurate examples of how to combat sound delay and did not demonstrate knowledge of the concept.
Overall Essay Composition	Response is composed exceptionally and contains more than the desired content of the essay. No grammatical or syntactical errors.	Response is well-composed and contains the desired amount of content. Few grammatical or syntactical errors.	Response is composed in partial error and is lacking the full amount of desired content. Several grammatical and syntactical errors take away from effectiveness of the essay.	Response is poorly composed and does not contain the desired content. More grammatical and syntactical errors present than not.

Appendix B

Demonstration of Calculation Processes for Central Tendency and Dispersion

Raw Test Scores
100 95.7 95.7 95.7 95.7 94 93 93 93 91 89 86 78.5 77 65

Mean
[100 + (95.7 x 4) + 94 + (93 x 3) + 91 + 89 + 86 + 78.5 + 77 + 65] / 15 = 89.49

Median
93

Mode
95.7

Range
100-65 = 35-point range

Variance (s^2)
s^2 = 86.66

$$s^2 = \Sigma \frac{(x - mean)^2}{n - 1}$$

x = each data point
mean = the sample mean
n = sample size

Data	**x – *mean***	**(*x* – *mean*)^2**
65	-24.29	599.76
77	-12.49	156
78.5	-10.99	120.78
86	-3.49	12.18
89	-0.49	0.24
91	1.51	2.28
93	3.51	12.32
93	3.51	12.32
93	3.51	12.32
94	4.51	20.34
95.7	6.21	38.56
95.7	6.21	38.56
95.7	6.21	38.56
95.7	6.21	38.56
100	10.51	110.46

All values in the third column summed together = 1,213.24.

1,213.24 / 15-1 = 86.66

Standard Deviation
8.99 points from the mean

$$\text{SD} = \frac{\sqrt{\Sigma(x-\mu)^2}}{N}$$

Step 1: find the mean. Mean = 89.49
Step 2: for each data point *x*, find the square root of its distance from the mean |*x*-μ|^2

$$\Sigma\,|x-\mu|^2 = \frac{1{,}213.24}{15} = 80.88$$

$$\sqrt{80.88} = 8.99$$

Item Difficulty
P = R/T
P = item difficulty index
R = number of correct responses
T = total number of responses

Average item difficulty for most difficult 50% of items:
[(0.933 x 9) + (0.867 x 8) + 0.8 + (0.67 x 2) + 0.6] / 21
18.073 / 21 = 0.86

Item Discrimination
Upper 33% of scores: 100, 95.7, 95.7, 95.7, 95.7
Middle 33% of scores: 94, 93, 93, 93, 91
Lower 33% of scores: 89, 86, 78.5, 77, 65

Step 1: Count the number of students in the higher group that got each item correct, for all items.
Step 2: Divide the number of students who got the item correct by the total number of students in the high group (in this case, divide by 5).
Step 3: Repeat steps 1 and 2 for the lower group.
Step 4: for each item, subtract the proportion of low scoring students who got each item correct from the proportion of high scoring students who got each item correct. The remainder is the discrimination index.

*See Appendix C and Appendix D for detailed findings on item difficulty and item discrimination.

Average discrimination index for most discriminating 50% of items:
[(0.2 x 14) + (04. x 5) + (0.6 x 3)] / 21
6.6 / 21 = 0.3

Appendix C

Item Difficulty and Discrimination Summary Table

Item Type	Item Number & Type	Item Difficulty	Item Discrimination (DI)
Matching	Question 1	1.0	0
	Question 2	1.0	0
	Question 3	1.0	0
	Question 4	0.933	0.2
	Question 5	0.867	0.2
	Question 6	1.0	0
	Question 7	0.867	0.2
	Question 8	1.0	0
	Question 9a	1.0	0
	Question 9b	0.933	0.2
	Question 9c	1.0	0
	Question 9d	1.0	0
	Question 9e	1.0	0
	Question 9f	0.8	0.6
	Question 9g	0.867	0.4
	Question 9h	0.933	0.2
Short Answer	Question 10	1.0	0
	Question 11	0.867	0.2
	Question 12	0.867	0.4
	Question 13	1.0	0
	Question 14	1.0	0
	Question 15	0.67	0.2
	Question 16	1.0	0
	Question 17	1.0	0.2
Multiple Choice	Question 18	1.0	0
	Question 19	0.867	0.2
	Question 20	0.933	0.2
	Question 21	1.0	0
	Question 22	0.933	0.2
	Question 23	0.933	0.4
	Question 24	0.67	0.6
	Question 25	1.0	0
	Question 26	1.0	0
	Question 27	1.0	0
True/False	Question 28	1.0	0
	Question 29	1.0	0
	Question 30	0.933	0.2
	Question 31	0.867	0.2
	Question 32	0.867	0.4
	Question 33	0.6	0.6
	Question 34	0.933	0.2
Essay	Question 35	0.933	0.4

Calculations for item discrimination can be found in Appendix D.

Appendix D

Table of Discrimination Index Calculations

Highly discriminating items (>0.3) are **bolded.**

Item Type	Item Number	High Scoring Proportion	Low Scoring Proportion	Discrimination Index
Matching	Question 1	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 2	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 3	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 4	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 5	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 6	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 7	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 8	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 9a	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 9b	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 9c	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 9d	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 9e	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 9f	**5/5 = 1**	**2/5 = 0.4**	**1 – 0.4 = 0.6**
	Question 9g	**5/5 = 1**	**3/5 = 0.6**	**1 – 0.6 = 0.4**
	Question 9h	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
Short Answer	Question 10	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 11	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 12	**5/5 = 1**	**3/5 = 0.6**	**1 – 0.6 = 0.4**
	Question 13	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 14	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 15	3/5 = 0.6	2/5 = 0.4	0.6 – 0.4 = 0.2
	Question 16	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 17	1.0	0.2	1 – 0.8 = 0.2
Multiple Choice	Question 18	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 19	4/5 = 0.8	3/5 = 0.6	0.8 – 0.6 = 0.2
	Question 20	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 21	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 22	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 23	**5/5 = 1**	**3/5 = 0.6**	**1 – 0.6 = 0.4**
	Question 24	**4/5 = 0.8**	**1/5 = 0.2**	**0.8 – 0.2 = 0.6**
	Question 25	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 26	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 27	5/5 = 1	5/5 = 1	1 – 1 = 0
True/False	Question 28	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 29	5/5 = 1	5/5 = 1	1 – 1 = 0
	Question 30	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 31	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
	Question 32	**5/5 = 1**	**3/5 = 0.6**	**1 – 0.6 = 0.4**
	Question 33	**4/5 = 0.8**	**1/5 = 0.2**	**0.8 = 0.2 = 0.6**
	Question 34	5/5 = 1	4/5 = 0.8	1 – 0.8 = 0.2
Essay	Question 35	5/5 = 1	3/5 = 0.8	1 – 0.8 = 0.4

Appendix E

Demonstration of Calculation Processes for Internal Consistency: Split-Half Reliability Using the Pearson Correlation Formula

$$r = \frac{\Sigma\,(x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma(x - \bar{x})^2 (y - \bar{y})^2}}$$

where
r = Pearson correlation coefficient
x = x-value in a sample
$\bar{x}$= mean of x-values in sample
y = y-value in a sample
$\bar{y}$= mean of y-values in sample

$\bar{x}$ = [(100 x 9) + 96.3 + 92.6 + (88.88 x 2) + 81.5 + 72.22] / 15 = **94.69**

$\bar{y}$ = [100 + 97.67 + 95.35 + (93.02 x 4) + 90.7 + (88.37 x 2) + 86.05 + 81.4 + 76.74 + 72.09 + 47.67] / 15 = **86.43**

$\Sigma(x\text{-}\bar{x})$ = **0.03**
$\Sigma(y\text{-}\bar{y})$ = **-1.88**

Students (total score)	x-score (#1-14)	$(x\text{-}\bar{x})$	y-score (#15-35)	$(y\text{-}\bar{y})$	$(x\text{-}\bar{x})$^2	$(y\text{-}\bar{y})$^2
1 (100%)	100	5.31	100	13.57	28.2	184.14
2 (95.7%)	100	5.31	93.02	6.59	28.2	43.43
3 (95.7%)	100	5.31	93.02	6.59	28.2	43.43
4 (95.7%)	100	5.31	93.02	6.59	28.2	43.43
5 (95.7%)	100	5.31	93.02	6.59	28.2	43.43
6 (94%)	88.88	-5.81	97.67	11.24	33.76	126.34
7 (93%)	100	5.31	88.37	1.94	28.2	3.76
8 (93%)	100	5.31	88.37	1.94	28.2	3.76
9 (93%)	96.3	1.61	90.7	4.27	2.59	18.23
10 (91%)	100	5.31	86.05	-0.38	28.2	0.14
11 (89%)	100	5.31	81.4	-5.03	28.2	25.3
12 (86%)	72.22	-22.47	95.35	8.92	504.9	79.57
13 (78.5%)	88.88	-5.81	72.09	-14.32	33.76	205.06
14 (77%)	81.5	-13.19	76.74	-9.69	173.98	93.9
15 (65%)	92.6	-2.09	47.67	-38.76	4.37	1,502.34

$$r = \frac{\Sigma\,(0.003)(-1.88)}{\sqrt{\Sigma(0.0009)(3.53)}} = \textbf{-0.10}$$