

Evaluating Student Work Using Models Derived from Those Used in Nationally Administered Examinations

Joel Phillips

Each year thousands of students take one of two music examinations administered nationally by the Educational Testing Service. Students who hope to attend graduate school in music take the revised Graduate Record Examination (GRE) in Music and high school students who wish to receive undergraduate credit for the first year of music theory take the Advanced Placement (AP) Examination in Music Theory.

For a number of years both of these examinations have incorporated tasks typical of those given in college music theory classrooms. As examples, students are asked to realize a figured bass, take melodic and harmonic dictation, or harmonize a melody. Because of the complexity of these tasks the responses must be judged by human experts. These experts must agree upon the way in which an item will be scored and apply those standards with such consistency that, given a particular student response, all experts who have been so trained should arrive at the same score within a very small margin of error.

In this article I will describe and illustrate the types of judgments made in these examinations. Then I shall demonstrate how I have applied models derived from the scoring guides of these examinations to meet my own classroom needs. Because the nature of the feedback on the national examinations differs from that found in the classroom, I will also demonstrate the type of feedback I give students, with particular emphasis on peer evaluation and collaborative learning. Because I have had the privilege and pleasure of training readers for each of these examinations I can offer what I hope to be an interesting perspective to the task.

Though these models of scoring might not be appropriate for every type of work one evaluates in one's own classroom, there are many possible derivations that have the potential to make day-to-day judgements faster, fairer, and more consistent. Among the people who might see immediate benefit from these examples are teachers of classes with large enrollments, departments that offer multiple sections of the same class but which have different teachers, persons who supervise graduate assistants, persons who must make real-time judgements (such as those made during auditions or in aural skills classes), and persons who score placement tests.

There are two fundamental types of judgements rendered in these examinations. These I will characterize as *analytic judgements* and *holistic judgements*. Some scoring guides are purely analytic and some guides are purely holistic. Sometimes the guides are comprised of a combination of both analytic and holistic judgements.

The analytic judgements are typically used for items that are indisputably right or wrong. The harmonic dictations given on either examination lend themselves well to this type of judgement. In these questions students are asked to record the soprano and bass pitches of a harmonic dictation and to supply the Roman numerals and figures of the progression they hear. From the judge's point of view the student responses are deemed either right or wrong. A pitch is either correctly notated or it is not. The chord choice a student circles is either correct or it is not. A second example of analytic judgement is the regular scoring guide for the AP melodic dictation. In the regular scoring guide each segment (usually half a measure) completely correct in both pitch and rhythm is awarded one point. After judging all segments one additional point is added to the subtotal to obtain the final score. (This last point helps discriminate people who did everything wrong (=0 points) from their colleagues who did at least one thing right (=1 point).)

Consider Example 1 to be a melodic dictation typical of that played in a first year class. Consider Example 2 to be a student response to the melodic dictation played in Example 1. This response features five correct segments each worth one point to which an additional point is added for a total score of six. Each incorrect segment is marked with an X.

EXAMPLE 1.



EXAMPLE 2.



Analytic judgements may also be used for items that have clearly defined, generally accepted answers. For example, when students are asked to resolve an augmented sixth chord, there are but a small number of possible responses that are characteristic of the common practice period. Readers are trained to anticipate those potential responses and to accept no others.

Example 3 illustrates a problem of this type followed by four acceptable answers. Note that there is some flexibility in this process; after all we are dealing with music! If a student paper reveals a musically satisfying resolution that is not on the list of acceptable responses, the judge simply asks the person who trains the readers to permit the unanticipated solution. If you are about to read several hundred papers, these types of relatively objective, analytic judgements can save a great deal of time.

Holistic judgements are used in situations where there are many possible correct solutions and for responses in which the answers might depend upon the context in which they are found. Thus in holistic judgements an entire example or a portion of an example might be viewed with a certain subjectivity. The purest of the holistic judgements might be characterized by the use of descriptive words or a phrase that applies to a response. Perhaps the most familiar type of holistic scale is the traditional grading scheme of A=excellent, B=good, C=satisfactory, D=unsatisfactory and F=failure.

In the composition question on the AP examination students are asked to write a bass line to counterpoint a given melody, and to write the harmonic implications of the line below the staff. A judge studies each phrase and decides upon a score using descriptive words or a phrase to describe the perceived quality of that phrase. For a phrase that demonstrates an "excellent treatment of the bass, harmonic progression, and Roman numerals" a judge awards 4 points. For a phrase that demonstrates a "good treatment of the bass, harmonic progression, and Roman numerals" a judge awards 3 points, and so on. As to what criteria constitute an "excellent treatment," the subjectivity of the judges is tempered in an extensive training lasting more than half a day that includes more than thirty samples from actual student papers as reference. An additional ho-

EXAMPLE 3.

Compose a resolution of the following chord in common practice style.

acceptable, likely solutions

The image displays five musical staves, each representing a different resolution of an A6 chord (A major with a sixth, F#) in C major. The staves are arranged vertically, with the first staff on the left and the fifth on the right. Each staff shows the A6 chord in the first measure, followed by a resolution in the second measure. The resolutions are: 1) A6 to D major (D4, E4, F#4, G4), 2) A6 to E minor (E4, F4, G4, A4), 3) A6 to D minor (D4, E4, F4, G4), 4) A6 to E major (E4, F#4, G4, A4), and 5) A6 to D major (D4, E4, F#4, G4). The staves are labeled A6, A6, A6, A6, and A6 from left to right. The notation includes treble and bass clefs, a key signature of one flat (Bb), and a common time signature (C). The notes are written as quarter notes.

listic judgement is made after each phrase is judged. The judgement reflects the musicality of the composition as a whole.

Example 4 is an actual student response to the composition question and includes the rationale for the decisions made in awarding the submission a score of 16. Note that in this student's paper an additional discretionary point is awarded for the overall musicality of the composition.

Another example of holistic judgement is the use of alternate scoring procedures for sight singing on the AP examination. These alternate procedures are used for responses in which students substantially disregard either the pitch or the rhythm of a melody. (For typical responses a regular scoring guide, similar to the analytic guide used to judge melodic dictation, is used. Consider the music of Examples 1 and 2 to be a sight-singing melody and a student's attempt to sing the melody.) If many of the pitches are incorrect in a sight-singing melody, two points may be awarded for rhythm that is correct or mostly correct. Similarly, if many of the pitches are incorrect in a sight-singing melody, one point may be awarded for rhythm that is about half right.

Consider Example 1 to be a melody to be read at sight. Example 5 notates a student response that would invoke the use of the alternate scoring guide. Because this student disregarded the pitch and sang about half of the rhythm correctly, this response would receive a score of one point. Each rhythmic error is marked with an arrow. On the other hand if a response's rhythm is mostly inaccurate, up to four points may be awarded for correctly sung pitches. Example 6 notates a response that lacks rhythm, but in which the pitches sung are mostly correct. This response would receive a score of three points. Each pitch error is marked with an arrow.

In addition to these alternate scoring guides there are further examples of holistic judgements made in the sight-singing examination. The scores of 0 and 1 can also be applied to a response with "no redeeming quality" and a response with "some redeeming qualities" respectively.

Sometimes analytic and holistic judgements are combined to obtain the score for a single item. In judging the free counterpoint item on the GRE each segment (usually one beat of music) of a response is worth one point. If there is any major error, such as paral-

EXAMPLE 4. Sample AP Composition Question



Chord Analysis: V^2 I^6 V_4^6 I V I V_4^6 I^6 I_4^6 ii ii^6 V^7



Chord Analysis: I ii I^6 V vi iii vi vii^0 I I^6 ii^6 V I

This is an example of an excellent response. The harmonic motion of the first phrase is completely logical, including the opening on the $V_4/2$ chord that resolves to the I^6 ; four points are awarded for this phrase. The second phrase is a bit rougher, due to the use of the passing $V_6/4$ on a strong beat (the downbeat of m.3) and the awkward I_4^6 on the last beat of that measure. If one looks to the third beat of m.4 for a cadence, the landing on the dominant works well. This is an example of “good treatment of bass, harmonic progression, and roman numerals,” resulting in 3 points. Phrase 3 works fine as is, though it could be improved by a passing vii^6 chord instead of the ii chord in m.5, and by a V/vi chord instead of the ii chord in m.6. It receives 4 points. Finally, the fourth phrase exhibits a logical bass line and harmonies (even though a $V_6/5$ would have been a better choice than a vii chord in root position for the pick up of this phrase), resulting in 4 points for the student. Thus, the student receives a total of 4 plus 3 plus 4 plus 4 points, resulting in 15 points. Because the bass line shows sensitivity to contrapuntal principles (contrary motion, combination of conjunct and disjunct motion, and so on), this student is awarded the extra “aesthetic” point, resulting in a final score of 16 points.

From p. 56 of the 1993 APTM Music Theory: Free Response Scoring Guide with Multiple Choice Section. © 1993 by College Entrance Examination Board. All rights reserved. Used with permission.

EXAMPLE 5.



EXAMPLE 6.



lel fifths, the entire point is subtracted; if there is a minor error, such as an unresolved leap, one half point is subtracted. There is a list of major and minor error types that accompanies the scoring guide to the problem. After making these analytic judgements the reader views the entire response holistically. An additional deduction may be made for responses that are especially unmusical.

Though there is far more to the scoring process on these examinations than has been presented here, the reader is now familiar enough with the concepts that I can demonstrate how I have adapted these procedures for my own use in class. I shall choose four models for the purpose of illustrating the process I use with the understanding that if one finds the ideas useful one will certainly be able to apply the concepts to one's own classroom. For each model to be discussed I asked myself four questions designed to help me structure the evaluation. They are:

1. What are the goals of the judgement?
2. What feedback, if any, is required?
3. What are the criteria for this evaluation? (The more unique the judgment, the more thorough I want to be and the more specific the criteria for evaluation need to be.)
4. Is there a way to get the evaluation from the students themselves (peer evaluation)?

The first model deals with scoring individual performances in sight singing. Many times a soloist and teacher are the only people who pay attention to a sight-singing performance. Others busily study the remaining melodies in case they might be called upon. I established several goals to improve this activity: 1) the teacher should not have the sole responsibility of engaging students in class activities, students must develop a sense of ownership in their music making; 2) performers should sing before an attentive audience of critical listeners; 3) each class activity must emphasize as many skill levels as possible; and 4) everyone should be involved during each performance.

I then established criteria by which the goals could be judged. I emphasize five skill areas in sight reading: pitch, rhythm, solfege, conducting and musicality. Each may be judged holistically, like each phrase in the AP composition question. For example, pitch may be judged *good to excellent*, *adequate to fair*, or *poor*, receiving scores of 2, 1 and 0 respectively. The five skill area scores are tallied. An additional, discretionary point may reward an outstanding job in four areas and a poor job in the fifth. For example, during an otherwise superior performance a performer who forgot to conduct might earn the extra point. To correct errors made in their initial performance, students may sing twice. The performer receives the higher of the two scores from each evaluator. The class practices this peer evaluation for several days before the first scores are recorded.

In order to maintain the intensity of interest I desire, I insist that everyone evaluate each performance, detecting errors and offering suggestions for improvement. Because they must evaluate the performers' conducting, the reviewers must watch the performers. Just like conductors, the reviewers must glance at their music, memorize a segment, look up at the performer and compare what they

hear with what they remember. These actions reinforce the skills I am attempting to develop in the students. Reviewers take this job very seriously because they know that their colleagues' grades depend on the accuracy of their evaluations. Moreover, their colleagues will potentially see these scores and comments.

After a performance I collect evaluations from two peers. If the two reviewers are within one point of each other's judgements, as they almost always are, I record the higher of the two scores. If there is a difference of two points, I record my score. Frequently there is complete agreement between both peer evaluators and myself. Example 7 is a handout I give my students when performing peer evaluations of sight reading.

This system of peer evaluation is derived from the scoring method used in the GRE. In the GRE most questions are read by two readers whose independent judgements must agree within one point of each other. On the rather infrequent occasions when there is a two point disagreement, a third, independent reading is made by the person who trained the item. This third score is recorded for the student.

This system of peer evaluation works beautifully for me. In addition to the scores recorded in each area of evaluation, student evaluators frequently offer fine suggestions for improvement or words of encouragement to their peers. My students are developing a much greater awareness for the art of critical listening. There is a focus in this type of class that I could never muster in the past. Best of all I have given up the role of "grim reaper" and assumed that of facilitator. For those of you who might wonder, given my example of an otherwise superior performance which lacks conducting, what the score of 9 really means, let me share with you that my class operates on a narrow scale; a 9 is a B+, an 8 a C, etc.

The second model is a harmonic dictation team project. For several weeks I use the beginning of each class to introduce many of the harmonic patterns inherent to common practice music. The class and I collectively derive "harmonic dictations" from the music we study by removing melodic diminutions from the surface of the music. Then we create new music through the systematic melodic diminution of these harmonic and melodic outlines. Each phrase we analyze is deconstructed into "chunks" which may be catego-

EXAMPLE 7: Peer Evaluation of Sight Singing Performance

Areas of Evaluation	Possible points*
Pitch	2, 1, 0
Rhythm	2, 1, 0
Solfege	2, 1, 0
Conducting	2, 1, 0
MUSICALITY	2, 1, 0
total	10 points

*2 means good to excellent; 1 means adequate to fair; 0 means poor, inadequate or absent

If you have scored the performance as noted above and the score seems too low or too high to you, you may adjust the final score by one point up or down. For example, if the score is an 8 because the conducting was poor, but everything else was outstanding, you may add 1 discretionary point for a final score of 9 points.

Each performance will be evaluated by everyone in the class. I will collect the evaluations from two of your peers. If the peers are within one point of each other, the recorded score will be the higher of the two. If the peer scores differ by two or more points, my score will be recorded.

Evaluation Form

Please copy or emulate this form for use when evaluating each other in class.

Evaluation of _____ Evaluator _____
 Date _____ Singing Exercise # _____

Areas of Evaluation	Possible points* First Attempt	Possible points* Second Attempt
Pitch	2, 1, 0	2, 1, 0
Rhythm	2, 1, 0	2, 1, 0
Solfege	2, 1, 0	2, 1, 0
Conducting	2, 1, 0	2, 1, 0
MUSICALITY	2, 1, 0	2, 1, 0
optional discretionary point (+ 1)	_____	_____
total score	_____	_____

*2 means good to excellent; 1 means adequate to fair; 0 means poor, inadequate or absent

Comments or suggestions for improvement:

ized as tonic establishment chunks, connective chunks, or cadential chunks. We record these chunks for later inclusion on a worksheet that accompanies the collaborative project guidelines. Thus each example students create serves two purposes: the music might be used as a harmonic dictation or as the outline for composing or improvising a phrase.

After students are familiar with the processes just described I sort them into teams. The teams are comprised of two partners each of whom composes, performs, teaches, listens to, and scores harmonic dictations. Unlike the classroom method, the speed and interaction between participants may be customized. Hints may be given and questions may be asked as long as no specific detail is revealed.

The projects are scored analytically and the method of judgement is conceptually identical to that used in each of the nationally administered examinations. The grade is determined by using a simple formula: students add the number of soprano pitches, bass pitches and Roman numbers/figures that are correct and divide that number by the total number of possible responses to obtain the score for the dictation.

Like the sight singing peer evaluation, there is a suggested format for the harmonic dictation peer evaluation. Once again the burden of responsibility rests with the students. Dictation becomes an active, musical process. The team activity eventually replaces the daily harmonic dictations in class. For additional practice these team projects are also supplemented with individual work using computer-assisted instruction software.

The third model deals with the judgement of sight singing in juried examinations. Each semester my colleagues and I must jury the examinations of all persons enrolled in our sightsinging courses. A minimum of two faculty members must hear each student—the student's teacher and a department member at-large. Sometimes the second faculty member can only stay for part of the examination and is relieved by a third member of the department.

I have two scoring schemes for this type of evaluation. My first scoring scheme is derived from the AP examination, which now includes an evaluation of sight singing. Each judge listens to the performance, awarding one point for each segment of music performed correctly in both pitch and rhythm in the case of the melo-

dies, and rhythm alone in the case of non-pitched rhythms. One point is added to the score if there are no hesitations or restarts in the performance. If a minor melody is consistently transformed to major during a performance, the performance is scored as if the tune had been written in the major mode. Three points are then deducted from this subtotal to obtain the final score. The second scheme is identical to the method discussed earlier for the peer evaluation of melodies.

Regardless of which scheme I use, the two teachers' scores are later compared by the teacher of the examinees who averages them. These scores are usually identical and seldom differ more than one point. This scoring method permits the performance of two melodies and two non-pitched rhythms during a seven-minute individual hearing. Because these examinations extend for hours this method of scoring helps the judges maintain their focus. If students have questions about their performance after the examination, there is a record of the exact places where they made their errors. To prepare for the experience of the jury I hold mock juries during the last two classes in order to minimize test anxiety.

There are some differences when compared to the judgements made in the AP examination. Like other items in the AP examination, the sight singing portion is scored by a single judge, though there are frequent checks of the readers' judgements by the item trainers to ensure continued vigilance on the part of the readers. The judge listens to a tape of a student performance until a judgement can be rendered—usually two or three times.

Many simpler models exist in my revised "bag of tricks" for the classroom. In the past I devoutly scored every homework assignment on a scale of 100 without really appreciating how much time I spent fitting the assignment to the scale and tallying the scores, but to what end? For most daily work I now use a system that minimizes the time I spend scoring a paper and maximizes the time I spend giving feedback on the paper. The system is an embarrassingly simple, holistic scale. Perhaps you recognize it? A *check* means the paper demonstrates a fair to excellent grasp of the material, a *minus* means the paper demonstrates a poor grasp of the assignment and a *zero* means the assignment was not turned in or was substantially incomplete.

EXAMPLE 9. Harmonic Dictation Team Project, p.1

This team project will replace much of our classroom activity in harmonic dictation. Because you will work privately with a peer there will be more flexibility than is possible in our classroom setting. This project will be an excellent opportunity to reinforce many diverse concepts because you will each play the role of composer, performer, listener and teacher.

Stage 1: Compose your harmonic dictations. The progressions you compose should be comprised of "chunks" from the HD Chunks handout (or from other pre-approved dictation materials). Choose first a tonic establishing chunk, then a cadential chunk and, finally, a chunk that links the tonic to the cadential chunk. I strongly encourage you to derive your progressions from the music we are studying. If so, please cite the piece, composer name and the measures from which you have derived your music. Progressions should be approximately 8-10 chords in length, written in SATB or 4-part keyboard style with good voice leading. If you have a more experienced partner you may, of course, compose longer progressions. Once you see the individual strengths and weaknesses of your partner, compose progressions that will best help your partner improve on his or her weaknesses.

Stage 2: Have the teacher review your compositions. You are required to submit your progressions to the teacher for approval prior to giving them to your partner.

Stage 3: Perform your music for your partner. At a private meeting with your partner you will perform your musical examples and, as a teacher, help your partner to improve her or his skills. Use a similar approach to that developed in class. You may, however, play the example more often than is possible in class as well as at different tempos. You may give your partner generic "hints," but do not reveal any specific information about the particular example being performed. Now that you understand melodic diminution, remember that these dictations are the outlines of real music and, as such, deserve to be performed as musically as possible.

Stage 4: Listen and take dictation. This stage is exactly what you are accustomed to from your class experience. However, now you have the opportunity to interact more freely with your teacher/partner. You may ask to hear the progressions at different tempos. Perhaps you may wish to hear the bass voice more loudly, etc. For obvious reasons, do not ask questions of your partner that would compromise your own learning experience. Don't use your partner as a crutch!

Stage 5: Report on the progress of your peer. Each partner must report on the progress of his or her peer by the specified deadlines below. You are required to document each meeting with your peer. If your partner has special needs, please feel free to consult with me for helpful hints.

EXAMPLE 9. p.2**Deadlines for Reporting Individual Progress:**

High Noon, Friday, 25 October
 High Noon, Friday, 8 November
 High Noon, Friday, 22 November
 High Noon, Wednesday, 11 December

You may use copies of this page to turn in as a progress report.

Report of Progress in Harmonic Dictation

Evaluation of _____ Evaluator _____

Date of Meeting: _____ Duration of meeting: _____ minutes

1. Attach a copy of the progressions played for your partner. Indicate the number of times you played each progression on your music.
2. Attach your partner's graded dictations. (To grade, add the number of soprano pitches, bass pitches and Roman numbers/figures that are correct and divide that number by the total number of possible responses to obtain the score for the dictation. For example, assume a dictation has 9 soprano notes, 8 bass notes and 7 chord symbols and that your peer misses 2 notes and one chord symbol: $21 \text{ correct} / 24 \text{ total responses} = 87.5$. Round up to 88%.)
3. Make a list of the hints you gave your partner.
4. Make a list of the suggestions you gave your partner for his or her improvement.

What I do with the additional time is to give much more feedback to the student. To do so I employ numerous time savers. When judging figured bass realizations, for example, I use a system in which each of the common errors is assigned a number. Objectionable parallels is given the number 1, improper use of the six-four chord the number 12, incorrect suspension the number 17, etc. The students have a "legend" that lists the errors by number, but which also refers them to an outline we built together in which those errors are explained in some detail. These errors are grouped into major errors and minor ones, as in the counterpoint of the GRE or the figured bass in either examination.

What probably seems obvious to the reader is the fact that these ideas are so simple. Yet they were not apparent to me until my experiences with these examinations. Being an ETS reader turned out to be better than a pedagogy seminar! I now realize that the scoring decisions I was making were not only unnecessarily laborious, but also somewhat capricious. Further, I was so accustomed to spending time devising scoring schemes for each assignment and scoring student papers that I was unable to give as much feedback as my students deserved. By combining my own methods of giving feedback with the modified scoring methods used on these examinations, I believe I now operate more fairly and efficiently as a teacher.

For those people who might be interested in further information pertaining to these examinations, the names and addresses of the test development specialists at ETS who are responsible for overseeing the development of these instruments are listed below. Both would be happy to provide you with additional details regarding the content and scoring of these examinations as well as information on the potential opportunities to participate as a reader in the scoring process.

Janet Palumbo, GRE Music
Educational Testing Service
Mail Stop 24-N
Princeton, NJ 08541
(609) 683-2917
jpalumbo@ets.org

Janet Waanders, AP Music Theory
Educational Testing Service
Mail Stop 22-N
Princeton, NJ 08541
(609) 683-2914
jwaanders@ets.org