The Evaluation and Design of an Undergraduate Music Theory Placement Exam

Barbara Murphy

According to the National Association of Schools of Music 1999-2000 Handbook, all music majors must have certain competencies in aural skills and analysis, composition and improvisation, repertory and history, technology, and synthesis of ideas.[1] Generally, students meet these requirements by taking courses in Basic Musicianship which include courses in Theory, Ear-training, and Musicology. These courses comprise 20-35 percent of their curriculum (24-42 hours of a typical 120 semester hour curriculum).[2] Even though NASM states that to be admitted to a baccalaureate degree in music, students should possess a musical background that allows them to "relate musical sound to notation and terminology both quickly and accurately enough to undertake basic musicianship studies,"[3] many students enter college without this prerequisite knowledge. Some students enter music study having had no theory background; their only exposure to music has been in band or choir where little or no explanation of theoretical concepts has taken place. Other students may be able to read multiple clefs and play certain scales and intervals but may not know what they are playing. Still other students have taken theory in school or summer camp and they may be knowledgeable in theoretical concepts through chords or even part-writing.

To evaluate entering students, most schools of music give placement exams on rudimentary music topics. Because existing stan-

---

[1]National Association of Schools of Music. 1999-2000 Handbook. (Reston, VA: National Association of Schools of Music, 1999), 78-79.

[2]Ibid., 81-91.

[3]Ibid., 72.

dardized exams do not measure the material covered in undergraduate music theory courses, most schools create in-house, or criterion-referenced, exams. Although these exams have content validity, very few evaluations have been conducted on their results to determine the reliability and predictive validity of the exams.

The purpose of this article is to discuss the steps taken to evaluate the Theory Placement Exam used at the University of Tennessee, Knoxville and the steps taken to create a new Music Theory Placement Exam that could be used to place incoming music majors and minors into Theory or Fundamentals of Music classes based on their knowledge and background in theory. The new exam consists of items in a SuperCard stack and is administered via the computer. This article describes an evaluation procedure that determined the extent to which the exam appropriately and consistently placed students in a music theory class; that is, the procedure used to determine the predictive validity and reliability of the Undergraduate Music Theory Placement Exam.

Related Research

In recent years, there has been much development of computer assisted instructional programs in music and in theory in particular. However, only a couple of placement programs have been written and evaluated. These programs are described below.

James P. Colman's Program

In 1990, James Peter Colman reported on his development and validation of a theory placement exam for use at Michigan State University.[4] His goals were (1) "to create a diagnostic test capable of measuring the current music theory achievement of incoming students so that statistical data could replace the assumptions made by college theory professors," (2) "to determine whether the newly created test could function as a predictive variable in the evalua-

[4]James Peter Colman, "The Development and Validation of a Computerized Diagnostic Test for the Prediction of Success in the First-Year Music Theory Sequence by Incoming Freshman at Michigan State University." (unpublished Ph.D. Dissertation, Michigan State University, 1990).

tion of future success in music theory," and (3) "to provide a musically specific test that would give advisors of college music majors another aid for proper advising discussions."[5]

He developed the test by defining content areas, clarifying the behavioral objectives, and creating 90 test items which he then submitted to several teachers of theory (at MSU) for clarification and measurement of the objectives.[6] The test included items on aural skills as well as written theory skills. Topics included scales (27% weight), pitch notation (20%), notes and rests (13%), intervals (13%), triads (13%), key signatures (8%) and time notation (6%).[7] The test was delivered on Macintosh computers via HyperCard. It was presented in a linear fashion, although students were able to pass on any item and come back to it at the end of the test. The program created an answer sheet for each student with the student's score, as well as a file of the results for statistical analysis. Fifty-nine incoming freshmen took the 50-minute test during registration. The mean score was 56.75 with a standard deviation of 11.97. Colman used a test-retest method for assessing reliability resulting in a reliability estimate of .86.[8]

In addition to the results of the placement test, Colman looked at the grades the students received in the three quarters of theory classes that year. The sample size, therefore, dwindled from 59 to 32 students who completed the entire year of theory classes. From the results, Colman noted that the students' grades in the theory classes improved over the three terms, a fact which he attributed to practice by the students or to attrition. He found that students' grades on the test and their term grades correlated well, with a decline in the correlation in Spring term. The lowest correlation was between the test and the aural skills class grade. This low correlation could be the result of the low number (10%) of test items dealing with aural skills. Colman noted that his most striking result was "the discovery that the theory test correlated more strongly with the less sensitive grade point scale than it did with percentage grade scale."[9]

---

[5]Ibid., 1-2.
[6]Ibid., 41.
[7]Ibid., 40.
[8]Ibid., 53.
[9]Ibid., 67.

The final result was that, although the items on the test predicted course grades fairly well, some test items needed to be revised or discarded. However, this step was apparently never taken.

Timothy Smith's Program

In 1994, Tim Smith reported on his development of an adaptive computer test, called Ready or Not (RON), that the School of Music at Ball State University is now using as part of its "enrollment management program."[10] Smith defines an adaptive computer test as one that "could omit items that are too hard or too easy, compare student performance with statistical norms, and generate problems algorithmically."[11] This type of test "interacts with the user in a nonlinear fashion based upon responses as they are integrated with, and interpreted by, data."[12]

The goal of such a test is "more efficient and accurate identification of students at risk," where "efficiency is a measure of the time needed to accomplish that goal and accuracy is a measure of the validity of the assessment."[13] Smith's idea was that, if students show a high level of mastery, they should not be required to take test items lower in the hierarchy of difficulty, and so their test would be shorter than other students. Therefore, he decided to design his test "beginning with items of high discrimination [so that] the system would more quickly detect a trend toward mastery or non-mastery. After a trend had been established, substitution of criterion-referenced items would more efficiently and more accurately diagnose competence."[14]

Smith borrowed an equation from Bayesian mathematics called the Sequential Probabilities Ratio (SPR) to develop his program. By using this equation, the computer could be set to recalculate the probability ratio after each item and determine whether the student is a master or non-master of the topic in question.

---

[10]Timothy A. Smith, "An ExSPRT System Approach to the Assessment of Students Needing Remediation in Music Theory," *Journal of Music Theory Pedagogy* 8 (1994):179-200.

[11]Ibid., 180.

[12]Ibid., 186.

[13]Ibid., 189.

[14]Ibid., 187.

The resulting computer program tested twenty-two learning objectives, both written and aural. It began by posing items with high discrimination values and proceeded to items with lower discrimination values. These values were determined by analyzing items on a test previously given to 141 students at various universities. The computer program "begins by selecting one of the five most discriminating objectives and generating random problems within that objective. As each student demonstrates mastery or non-mastery of the objective, RON advances to one of the top five remaining objectives."[15]

At first, the students completed the test very rapidly, sometimes in less than ten minutes, and after being tested on only two or three objectives. In the second year, RON was told to withhold its assessment until the student completed at least five objectives.[16] But still RON often detected a need for remediation within fifteen minutes. One aspect of RON not discussed much in this article is the makeup of the test items themselves. Smith talks of the objectives and their discrimination values, but, unlike Colman who discusses the items at great length, Smith discusses them very little.

The Current Study

The aim of this project was to create a exam that would assess student achievement and place students into the proper class as well as keep data for analysis. This exam would not adjust to student responses like Smith's. The aim here was similar to Colman's study with the exception that this research would include a revision of the exam and an analysis of this revision.

The exams

The placement exam used at the beginning of this study, subsequently referred to as Placement 96, was a revision of an exam that had been used since 1993, subsequently referred to as Placement 93. Both Place-

[15]Ibid., 193.
[16]Ibid., 194.

ment 93 and Placement 96 covered written theoretical topics only; no ear training topics were included on the exams.[17]

Placement 93 consisted of twenty items. There were two notation items; for these items, the students were asked to name the lines and spaces of treble and bass clefs by writing the name of notes under the notes on the staff. For the one rhythm item, the students were to place bar lines in an example of 4/4 meter. Four items on scales followed; for each, students identified the type of scale written (major, natural, harmonic or melodic minor). Seven items on intervals were included, for which students identified the quality of the written interval. (For all but one item, all three responses used the same interval size.) On each of the last six items, students were to identify the quality of a triad presented in both treble and bass clefs.

Placement 96 consisted of twenty-five items, five each in the categories of notation, rhythm, scales, intervals and triads. All items were multiple choice questions with four possible answers. The items on Placement 96 were much more heterogeneous; the items were designed to test many types of knowledge in each category. The five notation items included one item testing knowledge of anchor notes (i.e., Which note is middle C?), two items testing the students' ability to identify pitches in treble and bass clef, one item on notation (i.e., Which is the correct notation for a Eb eighth-note?), and one item on the meaning of accidentals. The items on rhythm included two on note values (i.e., a given note is equal to how many beats or what combination of notes), one item on the number of beats in a given time signature, and two items for which students completed a measure with a note or rest. The items on scales and key signatures included two items on scale identification, two items on key signature identification and one item on parallel and relative major and minor scales. The interval items included one item

[17]At the University of Tennessee, like many other schools, ear training and written theory are two different classes. Most of the ear training at UT is delivered and tested via computer using a program called the Curriculum for Aural Training (C.A.T.) written in-house by Don Pederson and Mark Boling. Since these programs have their own placement exam which has been used successfully for several years, the placement exams used in this study covered only written theory topics.

on interval quality, one on interval size, one item on both quality and size, and two items for which the student chose which interval was the one specified. The five triad items included four on triad quality identification and one item on inversions.

Evaluation of Placement exams

Before any revision could be undertaken, the exams were evaluated to determine how they were performing as indicators of student knowledge and student success in theory classes. In addition, each exam item was evaluated to determine if it was separating those who knew the information from those who didn't (i.e., an item analysis). For these analyses, a macro entitled "Item Analysis for Multiple Choice Tests" from the SAS Sample Library was used. The macro computes descriptive statistics for analyses of data from multiple-choice exams. The macro also provides an item analysis. For the analyses, each observation (i.e., line of data) contains the answers of one Subject (S) to the exam items. The descriptive statistics resulting from the analysis on Placement 93 are shown in Table 1.

The statistics of most importance here are the mean score, the standard deviation, the KuderRichardson 20 (KR-20), and the Spearman-Brown Prophecy number. The mean score is the average score. A standard deviation is ". . . related to the distances . . . of scores . . . from their mean. . . . It is a measure of variability or dispersion such that a distance from one standard deviation below

Table I: Descriptive Statistics for Freshman Placement Exam 1993

Number of Items	20
Number of Subjects	115
Mean score	11.383
Variance of Scores	26.537
Standard Deviation	5.151
Kuder-Richardson 20	0.880
Standard error (from KR20)	1.783
Spearman-Brown Prophecy (from KR-21): To obtain a reliability of 90 the test should contain this many items	31

the mean to one standard deviation above the mean includes about 2/3 of the cases in most distributions of scores."[18]

Kurtz and Mayo explain the Spearman-Brown Prophecy formula as one that depicts the relationship that the "correlation between scores on one test and an alternate form of it will then be high."[19] Further, they explain that in deriving the formula, "it is assumed additional forms of the test exist or could exist that (1) all have equal standard deviations and that (2) all correlate equally with each other. In practice, it seems not to matter much whether or not these assumptions hold."[20] In addition, the formula can be used, as it was in this study, to determine how much an exam would need to be lengthened to achieve a specified reliability.[21]

To explain the Kuder-Richardson 20 formula, Kurtz and Mayo state that "Some time ago, a common procedure for determining the reliability coefficient was to split a test into comparable halves, to correlate scores on these halves, and then by the Spearman-Brown Prophecy formula . . . to estimate what the correlation would be between whole tests rather than between half tests. It was recognized that different ways of splitting a test into halves gave different results, but nobody did anything about it until Kuder and Richardson independently saw how to solve the problem. . . ."[22]

The Kuder-Richardson 20 formula solves the problem by making the following assumptions: " . . . Each [test] item measures exactly the same composite of factors as does every other item."[23] Secondly, it assumes that " . . . all inter-item correlations are equal."[24] Lastly, it assumes that " . . . all items have the same squared standard deviation or variance. . . ."[25] The resulting formula, the KR-20, "gives a good approximation to the values computed by the more rigorous formulas [the K-R (8) and K-R (14)]. . . ."[26]

---

[18]Albert K. Kurtz and Samuel T. Mayo. *Statistical Methods in Education and Psychology.* (New York: Springer-Verlag, 1979), 47.

[19]Ibid., 271.

[20]Ibid.

[21]Ibid., 272.

[22]Ibid., 265-266.

[23]Ibid., 267.

[24]Ibid., 268.

[25]Ibid.

[26]Ibid.

As shown in Table 1, 115 Ss took Placement 93 and averaged 11.383 items correct (56.9%) with a standard deviation of 5.151. The Kuder-Richardson 20 (KR-20) result was 0.880. The Spearman-Brown Prophecy result indicates that to obtain reliable results at the 0.90 level of reliability the test should contain at least 31 items instead of the 20 items on the exam.

The same analyses (i.e., descriptive statistics and item analysis) were also conducted on the subtests for notation, scales, interval quality and triad quality. The analysis was not run on the subtest for rhythm, since this subtest consisted of only one item. The descriptive statistics from this analysis are shown in Table 2.

Table 2: Descriptive Statistics for the Subtests of Freshman Placement Exam 1993

Subtest	1-Notation	2-Scales	3-Intervals	4-Triads
Number of Items	2	4	7	6
Number of Subjects	115	115	115	115
Mean score	1.8	1.983	3.765	2.965
Variance of Scores	0.214	1.859	4.479	4.385
Standard Deviation	0.463	1.364	2.116	2.094
Kuder-Richardson 20	0.383	0.644	0.74	0.80
Standard Error (from KR20)	0.363	0.813	1.079	0.936

The fact that the KR-20 result (0.80) is higher on Subtest 4-Triads than all the other subtests may be the result of the homogeneous nature of the relatively large number of items in this group (6). All items on Subtest 4-Triads dealt with triad quality and all items had the same four answer choices. Likewise, the low KR-20 result of Subtest 1-Notation (0.383) may be the result of its having only two fairly different items in this category. The conclusion that the homogenous nature of the test items has an effect on the score is supported by Anastasi in her description of Kuder-Richardson reliability as a: "method for finding reliability, . . . utilizing a single administration of a single form, [that] is based on the consistency of subject's responses to all items in the test. This interitem consis-

tency is influenced by two sources of error variance: 1) content sampling (as in alternateform and split-half reliability); and 2) heterogeneity of the behavior domain sample. The more homogenous the domain, the higher the interitem consistency. For example, if one test includes only multiplication items, while another comprises addition, subtraction, multiplication, and division items, the former test will probably show more interitem consistency than the latter. . . . It is apparent that test scores will be less ambiguous when derived from relatively homogeneous test."[27]

The statistics for Subtest 3-Intervals also follow the above assumption that the homogeneous nature of the items results in higher scores. The interval subtest had seven items—the largest number of items in a subtest—and all items tested interval quality; size was the same in each answer in all but one instance. The KR20 result is .74, the second highest, again indicating that the more homogenous the items and answer choices, the more reliable the exam is in indicating student knowledge.

Next, analyses were conducted on the answers from Placement 96. The descriptive statistics for the exam as a whole and the five subtests are shown in Table 3.

Table 3: Descriptive Statistics for Placement 1996

	Whole	Subtest 1 Notation	Subtest 2 Rhythm	Subtest 3 Scales	Subtest 4 Intervals	Subtest 5 Triads
Number of Items	25	5	5	5	5	5
Number of Subjects	90	90	90	90	90	90
Mean score	15.811	3.722	3.833	2.833	2.633	2.789
Variance of Scores	28.874	1.237	1.713	2.096	2.100	2.550
Standard Deviation	5.373	1.112	1.309	1.448	1.449	1.597
Kuder-Richardson 20	0.862	0.471	0.611	0.631	0.582	0.677
Standard Error (from KR20)	1.995	0.809	0.816	0.880	0.937	0.908
Spearman-Brown Prophecy (from KR-21): To obtain a reliability of .90 the test should contain this many	47					

[27]Anne Anastasi. *Psychological Testing*, 3rd ed. (New York: MacMillan Publishing Co, 1968), 84.

Although the mean score on the exam, 63% (15.811 of 25 items correct), is not an encouraging percentage, the exam does seem to be fairly reliable as evidenced by the Kuder-Richardson 20 result (.862). The KR-20 result on this exam is lower than the previous exam (.862 versus. 0.880) but not by much (0.18). It can, therefore, be assumed that both exams are fairly reliable and are testing at the same level. However, the number of items needed on this exam for a reliability level of 0.90 (47) is much higher than the 31 of the previous exam. This is most likely the result of the heterogeneous nature of the items on this exam versus the previous exam. This conclusion holds since on Subtest 5-Triads four of the five items test the same concept and this subtest has the highest KR-20 result (.67). This assumption is again borne out by the KR-20 results on Subtests 3 and 2 (.63 and .61 respectively). In each of these subtests, few different topics were covered; both Subtest 2 and Subtest 3 test only three different topics. The items on Subtest 1-Notation are the most diverse and, consequently, the KR-20 result on this Subtest is the lowest (0.471).

Revision of the Music Theory Placement Exam

After completing the item analyses, Placement 96 was revised. The number of items in each subtest was increased to ten, therefore increasing the number of items on the exam to fifty, following the guidelines presented by the Spearman-Brown prophecy. The statistics for each item of Placement 96 (i.e., the item analysis) were studied to determine which items needed to be changed or replaced. Items that separated the Ss in the top third of the group from those in the bottom third were kept and the other items were dropped or revised. Five similar items were added to each subtest. An attempt was made to make the items in each area more homogeneous. The resulting exam is now known as the UT Music Theory Placement Exam (or Placement 97). The general topic areas remained the same; specific topic areas included on the exam are listed in Table 4. An example item from each specific topic area is shown in Appendix 1.

As can been seen in Table 4, the items in the sections on Intervals and Triads were more homogenous—the items test the same information (i.e., eight items on triad quality) in the same format

Table 4: Topics Covered with Placement 97 exam items

General Topic Area	Specific Topic	Num.of items
Notation	Anchor Notes (middle C; A440)	2
	Clefs	4
	Enharmonic Notes	1
	Aspects of Notation (flags, accidentals placement)	3
Rhythm	Note Values	2
	Note/Rest Values - Simple time	3
	Note/Rest values - Compound time	3
	beats in simple time	1
	beats in compound time	1
Scales/KeySignatures	Types of scales	3
	Modes	1
	Key signatures	4
	Relative major and minor	1
	scale degree names	1
Intervals	Intervals - Quality only	2
	Intervals - Quality and size	4
	Identification of intervals	4
Triads	Triad - Quality only	8
	Triad Inversion	2

(i.e., all items have the same four possible answers in the same order). Items in other areas, such as notation, test many more aspects of the topics; these items are more heterogenous.

This new exam was entered into a SuperCard stack for delivery on Macintosh computers. The Supercard program consists of three windows: an introduction and instruction window, a exam item window, and a results window. The program works as follows from the students' perspective: The students see the introduction screen and, upon clicking the Continue button, are asked a series of questions about their name, social security number, major, and primary instrument. A file is created for each student to store this information as well as the answer to each item and the results on the exam. The exam then begins. The exam is presented in a linear fashion; all

students are presented every item in the same order. After the students choose an answer, a dialog box indicates whether the answer is correct or not, and the answer is recorded. Although telling the students that their answer is correct or incorrect may lead to some learning, there is probably little learning taking place since no explanation of why the answer was wrong is given.

At the end of the exam, the students are shown their results in the form of the percentage correct for each subtest and for the exam as a whole. In addition, the program indicates which class the students are recommended to take. If the students score at least 60% on each subtest and on the exam as a whole, they are recommended to take Music Theory 110B (a regular section of Theory I). If the students score lower than 60% on any subtest or the entire exam, they are recommended to take Music Theory 110A (a remedial section of Theory I) if they are music majors or Music General 100 (Fundamentals) if they are not music majors. Students may then print the results and take their recommendation with them. The scores and recommendations are stored in a computer file. A program is available to format and print the names of all the students who have taken the exam, their social security numbers, and their results for distribution to advisors and for their student files.

Placement 97 was given to students already enrolled in Music Theory 110A in the Fall of 1997 as a trial and subsequently to all students auditioning for the School of Music in the Spring of 1998 (119 students). An analysis was run on all 133 subjects' answers. The results are presented in Table 5.

The results of the analysis shows that the mean score on this exam was 32 (64.5%) with a standard deviation of 9.551. The KR-20 result for the exam was .909, the highest KuderRichardson result of all three placement exams. The Spearman-Brown prophecy number of 55 indicates that only five more items need to be added to this exam to obtain a reliability of .90, a great change over the almost doubling of the number of items on the last exam.[28] This new placement exam is thus an improvement over the older exams.

---

[28]In Spring 1999, five more items were added to the Placement 96 exam, one item in each of the five categories, making a total of 55 items on the exam. This new exam was given to 159 prospective music majors on their audition days. The KR-20 for this new exam indicated a reliability of .923.

Table 5: Descriptive Statistics for Placement 1997

	Whole	Subtest 1 Notation	Subtest 2 Rhythm	Subtest 3 Scales	Subtest 4 Intervals	Subtest 5 Triads
Number of Items	50	10	10	10	10	10
Number of Subjects	133	133	133	133	133	133
Mean score	32.241	7.692	7.233	6.248	5.271	5.797
Variance of Scores	91.230	2.927	4.286	5.703	6.760	9.284
Standard Deviation	9.551	1.711	2.070	2.388	2.600	3.047
Kuder-Richardson 20	.909	.585	.639	.726	.714	.827
Standard Error (from KR20)	2.884	1.103	1.243	1.251	1.392	1.267
Spearman-Brown Prophecy (from KR-21): To obtain a reliability of .90 the test should contain this many	55					

The KR-20 results on the subtests are also encouraging, since each one is an improvement over the KR-20 subtest scores on Placement 96. It is also interesting to note that the KR-20 scores increase with each subsection of the exam, again suggesting that the more homogenous the items, the higher the score.

Although the KR-20 scores change for the better on each subsequent subtest, the mean scores on the subtests fall from a high of 7.692 (76%) on notation to a low of 5.271 (52%) on intervals. (The mean score of the Triad subtest, 5.797, is higher than the intervals subtest but lower than all other subtests.) These mean scores are similar to those of Placement 96 (see Table 6), indicating that the students' knowledge has not differed much from one year to the next. All mean scores improved except for the score on the notation subtest. It is not yet certain why this decline occurred, but the change of score is minimal.

Correlation of Tests Scores to Grades

A very important factor in the evaluation process is to determine how well the exams serve as indicators of the students' performance in the theory classes they were recommended to take. To investi-

Table 6: Comparison of Mean Scores on Placement 96 and Placement 97

	Whole	Subtest 1 Notation	Subtest 2 Rhythm	Subtest 3 Scales	Subtest 4 Intervals	Subtest 5 Triads
Placement 96 Mean Scores	63.2%	74.4%	76.6%	56.6%	52.6%	55.7%
Placement 97 Mean Scores	64.4%	76.9%	72.3%	62.4%	52.7%	57.9%
Difference	+1.2%	+2.5%	-4.3%	+5.8%	+.01%	+2.2%

gate the relationship of exam scores to performance, the total correct scores on Placement 96 and Placement 97 were correlated to the grades the students received in their courses. Correlation, as Ott explains, is "One measure of the strength of the relationship between two variables x and y .... Given pairs of observations..., we can compute the sample correlation coefficient r .... r lies between -1 and +1. $r > 0$ indicates a positive linear relationship and $r < 0$ a negative linear relationship between x and y. $r = 0$ indicates no linear relationship between x and y."[29] What is hoped for, therefore, is a positive correlation coefficient close to 1.

To calculate the correlation coefficients, the students' grades for their theory classes had to be determined. For Placement 96, there were 69 students whose grades were found. Since Placement 97 has been given only one year, only 14 students had grades. It is important to remember that these 14 students were already taking theory classes and had been presented information on the exam prior to their taking the placement exam; their scores should thus be higher and correlate better with their grades.

The correlation of the total correct scores on the exam to grade point (e.g., 4.0 for A, 3.5 for B+) for Placement 96 was .4089 with a significant probability of .0005. Although this result is greater than 0 and therefore indicates a positive relationship, the relatively low score of .4089 indicates that the relationship is minimal. The same correlation for the 14 students taking Placement 97 was .7821 with a significant probability of .0009. Such a high correlation coefficient indicates that the exam better predicts students' grades in theory class and makes the exam much more useful for both the student and the teacher.

---

[29]Lyman Ott. *An Introduction to Statistical Methods and Data Analysis.* (North Scituate, Mass: Duxbury Press, 1977), 219-220

Conclusions and Future Plans

After studying the results of the analyses of the various exam scores and the correlation of exam scores to grades, one can conclude that the new exam, Placement 97, is an improvement over the older exams. One can also conclude that this improvement is due to the higher number of items on the exam and the homogeneous nature of the items on the subtests. Although a large amount of work has been completed to this date, more remains to be done in the areas of evaluation, revision, and programming projects for both the Placement exam and theory classes.

Once the students who took Placement 97 have completed their theory courses, their grades will be correlated to the total correct score on Placement 97. From the results of these correlations and the item analyses reported above, Placement 97 will be revised, improved and added to. The goal is to create a bank of exam items that the computer program can draw from so that students will not get the same items each time they take the exam. This process of testing, evaluation, and revision will continue for many years in order to create such a bank.

In addition, several improvements will be made to the exam. First, the items will be randomized, so that each student will receive the total number of items but not in the same order. Randomizing will help to prevent cheating as well as learning by the students since exam items on the same topic may not be given together. Other improvements to Placement 97 include improvements to the graphics and the reporting programs and the addition of a sample exam to be created for the Web. The sample exam would include items similar to those on the actual test so that students will know what to expect. Finally, mastery based tutorial programs on topics covered by the Placement exam are being created.

It is not the author's intent to make this exam a computer adaptive exam such as Tim Smith's "Ready or Not" for two reasons. First, the Placement 97 exam is performing well; its predictive validity is high (.78 as compared to Smith's RON at .52). Second, the author would like to control the minimum number of items on the exam and the exam makeup. The statistics used in this study (specifically the Spearman-Brown prophecy) indicate that at least 55 items

are needed to make the exam reliable. A computer adaptive exam may not give the students the minimum number of items; it may make its decision based on a smaller sample of items. Smith reported that he found this to be true; in the second year of testing of his program, he instructed RON to withhold a recommendation until it had tested at least five objectives. The new placement exam should contain at least 55 items covering the five topic areas on the exam.

The second aspect the author would like to control is the exam makeup. Smith's adaptive exam begins by choosing items in areas that are high in discrimination value. As a student demonstrates mastery or non-mastery, the student is moved on to another objective. This type of item generation seems to assume that students who do not do well on one topic (such as intervals) will not do well on topics that rely on this information (such as triads). This is not always the case. While analyzing the results of the exams in this study, it was found that some students score well on topics seeming to require more knowledge (such as triads) even though their scores on "easier" topics (e.g., intervals, rhythm) were low. The new placement exam should test the student on all areas and make a recommendation on their cumulative knowledge.

Once all this work has been completed, an entire package will exist for theory students similar to the many computer packages that exist for ear-training students. This package will consist of a exam of theory skills and reports of these results for instructors and administrators. The package will also include a system to route students to lessons on which they need more work. Reports of the students' results on the lessons will be available to teachers for the purpose of student evaluation and further research. It is hoped that this package of programs will be commercially available once it is completed. By having such a package, student skills may improve to the point that instructors can spend less class time on rudiments and more class time on issues of analysis and synthesis—the theoretical part of Theory class.

SELECTED BIBLIOGRAPHY

Anastasi, Anne. *Psychological Testing*, 3rd ed. New York: Macmillan Publishing Co., 1968

Colman, Peter James. "The Development and Validation of a Computerized Diagnostic Test for the Prediction of Success in the First-Year Music Theory Sequence by Incoming Freshman at Michigan State University." Unpublished Ph.D. dissertation, Michigan State University, 1990.

Kurtz, Albert K. And Samuel T. Mayo. *Statistical Methods in Education and Psychology*. New York: Springer-Verlag, 1979.

National Association of Schools of Music. *1999-2000 Handbook*. Reston, VA: National Association of Schools of Music, 1999.

Ott, Lyman. *An Introduction to Statistical Methods and Data Analysis*. North Scituate, Mass: Duxbury Press, 1977.

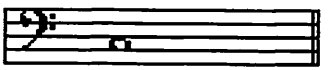
Smith, Timothy A. "An ExSPRT Systems Approach to the Assessment of Students Needing Remediation in Music Theory." *Journal of Music Theory Pedagogy* 8 (1994), 179-200.

APPENDIX 1:
SAMPLE ITEMS FROM PLACEMENT 97

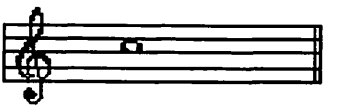
Notation

Topic: Anchor Notes
Which of the following is Middle C?

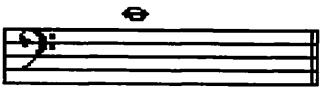
a.



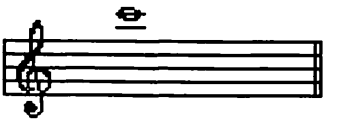
c.



b.

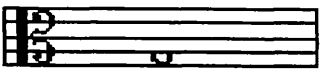


d.



Topic: Clefs
Which note is a D?

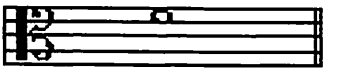
a.



c.

b.

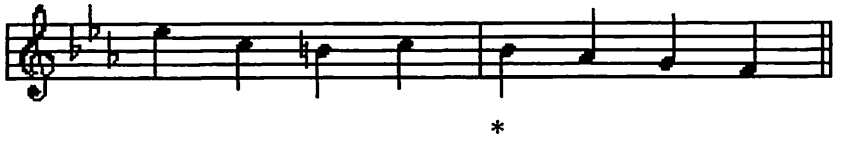
d.



Topic: Enharmonic Notes
Which pitch would sound the same as a C?

a. Db
b. Bx
c. B
d. Dbb

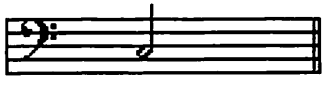
Topic: Aspects of notation
The note marked with an asterisk (*) below is a :



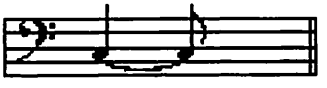
a. Bb
b. B
c. A#
d. D

Rhythm

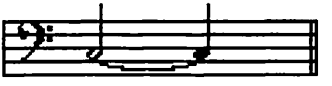
Topic: Note values
Which of the following is equal to the note below?



a.



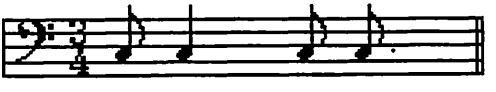
c.



b.

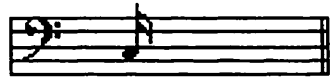
d.

Topic: Note/rest values - simple/compound time
Which note completes the following measure?



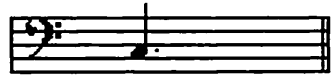
a.

c.

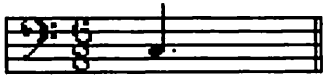


b.

d.



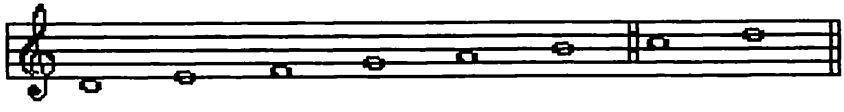
Topic: beats in simple/compound time
The note below would get how many beats?



a. 1/2 beat
b. 1 beat
c. 1 1/2 beats
d. 2 beats

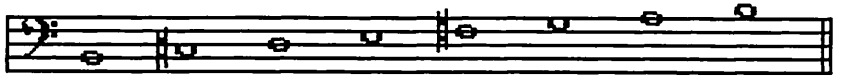
Scales/Key Signatures

Topic: types of scales
What scale is shown below?



a. D major
b. D natural minor
c. D harmonic minor
d. D melodic minor

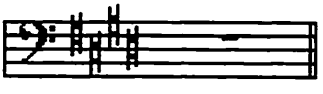
Topic: Modes
What scale is shown below?



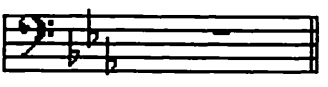
a. B Ionian
b. B Dorian
c. B Lydian
d. B Aeolian

Topic: Key Signatures
The key signature for E minor is which of the following?

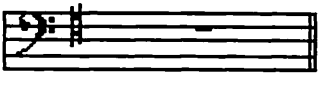
a.



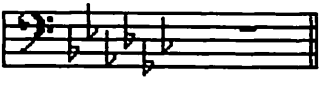
c.



b.



d.



Topic: relative major and minor scales
The relative major to A minor is:

a. F major
b. F# major
c. C major
d. C# major

Topic: Scale degree names
The fifth scale degree is called the:

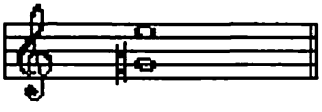
a. Tonic
b. Subdominant
c. Leading Tone
d. Dominant

Intervals

Topic: Interval quality
The quality of the interval shown below is which of the following?

a. Major
b. Minor
c. Perfect
d. Augmented

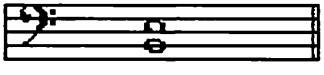
Topic: Interval quality and size
The interval shown below is which of the following?



a. Major 3
b. Minor 6
c. Augmented 5
d. Diminished 4

Topic: identification of intervals
Which of the following intervals is a Perfect 4?

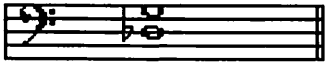
a.



c.

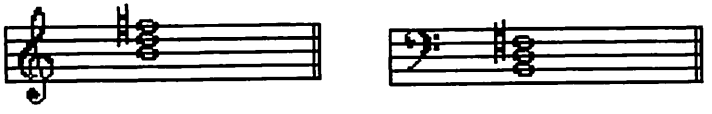
b.

d.



Triads

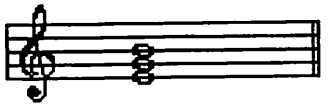
Topic: Triad quality
What is the quality of the following triad (shown twice in different clefs)?



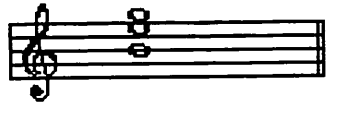
a. Major
b. Minor
c. Diminished
d. Augmented

Topic: Triad inversion
Which of the following chords is in first inversion?

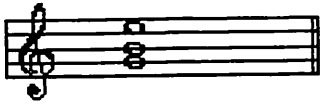
a.



c.



b.



d.