Journal of Music Theory Pedagogy

Volume 8 Article 8

1-1-1994

An ExSPRT Systems Approach to the Assessment of Students Needing Remediation in Music Theory

Timothy A. Smith

An ExSPRT Systems Approach to the Assessment of Students Needing Remediation in Music Theory

Timothy A. Smith

The typical freshman class in music theory is composed of students with formidable disparities in competence. It is not unusual for upwards of twenty percent of these people to withdraw or fail the first semester. Some failure is unavoidable, but excessive rates contribute to an inefficient use of faculty loads, student time, and money. It is possible to reduce the dropout rate if students with deficiencies are remediated.

It falls to schools of music to provide the instruments that will identify and advise unprepared students. No standardized test exists, however, for this particular diagnostic objective and age group. While the Aliferis-Stecklein Music Achievement Test does measure theory skills, its specifications also include history, literature, and performance. Further, both the Graduate Record Exam and the Aliferis Senior Comprehensive and Entering Graduate level test are too advanced to identify undergraduate students at risk. As a consequence, music schools often use instruments produced locally without an empirical basis for the interpretation of scores.

A systematic process of remediation requires that all students be tested, evaluated, and advised before registration. The logistical impediments of paper-and-pencil technology include scheduling appointments, proctoring and grading exams, and communicating results. By contrast, one might administer a computer test accessible on demand, with automated prognosis and recommendation available immediately upon completion. An *adaptive* computer test could accomplish the assessment even more efficiently and more

accurately. The computer could omit items that are too hard or too easy, compare student performance with statistical norms, and generate problems algorithmically.

This article documents a project undertaken by Ball State University to design and implement such a test. In late 1991 the project culminated in a computer adaptive test—called Ready or Not—that the School of Music has used since that date as part of its enrollment management program.[1] The article begins with a look at traditional methods of assessment, continues with a description of Ball State's pilot study of 1990-'91, and concludes with an explanation of the ExSPRT system now in use.

Related Research

The Harrison Study

Several researchers have compared student performance in the major with high school GPA, achievement tests such as the SAT, College BASE, or CAT, or music aptitude tests such as the College Music Aptitude Profile. Carole S. Harrison, for example, compared student performance in theory with three measures of achievement or aptitude (along with experiential measures such as the number of instruments played and years of private study).[2] She discovered the strongest correlations among theory grades and SAT math scores, high school GPA, and years of piano study. Surprisingly, Harrison found the music aptitude test to be a less than reliable measure when it came to predicting success in music theory. There is thus a need for a reliable theory-specific instrument that correlates with performance in the sequence.

---

[1]Before being admitted into the program students are required to submit SAT scores, audition, interview, write a short essay, and take the Ready or Not test.

[2]Carole S. Harrison, "Predicting Music Theory Grades: The Relative Efficiency of Academic Ability, Music Experience, and Musical Aptitude," *Journal of Research in Music Education* 38/2 (1990): 124-137. See also Harrison, "Relationship Between Grades in the Components of Freshman Music Theory and Selected Background Variables," *J.R.M.E.* 38/3 (1990): 175-186.

The Colman Study

In 1988 Peter Colman attempted to construct such an instrument.[3] Colman did not assemble an objective standard for the interpretation of his test or for the advising of students, but limited his work to constructing and validating a prototype that could be administered by computer. Colman arranged preskills into a behavior and content matrix out of which his panel of experts developed a table of specifications and ninety multiple choice items in the following proportions: 25% scales, 20% pitch notation, 15% notes and rests, 15% intervals, 15% triads, 10% key signatures, 5% time notation.

Colman scripted his test into a Hypercard stack which he used to assess 59 students at Michigan State University during registration week of the fall semester, 1988. Students averaged 50 minutes to complete the assessment. Colman selected twenty individuals at random to retake the test one week later, enabling him to establish the reliability of the test with a strong correlation of 86% and a small number of negatively discriminating items. Colman's longitudinal comparisons correlated these entering test scores with three grades for each student (aural lab, final grade in percentage points, final grade in a four-point grade scale) over a three-term span.

Colman's pretest scores correlated most highly with the four-point scale as averaged over the three terms—a correlation of 51%. The lowest correlation was between the test and the ear-training lab grades—32%. Colman was not surprised with this result because his test included only eleven aural tasks. Colman concluded that the test contained no major design flaws and that it could accurately diagnose a student's preparedness to enter college level theory in universities like Michigan State.

Colman started the study with 59 students and concluded with 32. Unfortunately, his sample was insufficient to generalize results to a larger population. Item response theory dictates that item pa-

---

[3]Peter Colman, "Development and Validation of a Computerized Diagnostic Test for the Prediction of Success in the First-Year Music Theory Sequence" (Ph.D. diss., Michigan State University, 1990).

rameter information be obtained from a pool of 200 to 1000 individuals.[4] Colman recognized this when he wrote: "An inherent problem . . . is the small size of the available sample and the great impact upon the study of students dropping out of the course" (p. 12).

The Ball State Pilot Study

In 1990, under joint auspices of Ball State University's Office of Research and Sponsored Programs and College of Fine Arts, I undertook to design an intelligent system for the purpose of assessing the skills of college freshmen beginning a course of study in music theory and identifying students at risk. The first phase of the research involved a Hypercard questionnaire, similar to Colman's, administered to volunteer freshmen at Ball State University, Indiana University, Biola University, Taylor University, and the University of Oregon[5] during the first week of the 1990 fall semester. From the nearly 200 students who took the pretest, I obtained complete linear data, including term grades, for 141. While this number is not sufficient to determine mastery on the basis of item response theory, it is more than enough to construct probabilities using Bayesian sequential probabilities ratios.

Besides gathering demographic, experiential, and attitudinal data, the questionnaire presented a series of aural, keyboard, and written tasks, recording the student's solutions and response times. The environmental data comprise the focus of ongoing analysis that may yet yield correlations of the type produced in the Harrison study. It is, however, the theoretical tasks that comprise the mathematical basis for the intelligent system described here.

The study began with an hypothesis that time spent on task would be a measure of fluency, that students who were masters would complete their tasks in less time than would nonmasters.

---

[4]Theodore Frick, "A Comparison of an Expert Systems Approach to Computerized Adaptive Testing and an Item Response Theory Model," paper presented at the Annual Conference of the Association for Educational Communications and Technology, February 1991.

[5]Thanks for the collaboration of colleagues Allen Winold, Edwin Childs, Fred Schultz, and Robert Hurwitz at the latter four institutions.

While the data supported this theory somewhat, I discovered that students who perceived themselves to be nonmasters often spent less time on tasks than did masters. It soon became apparent that while speed is, for some students, a measure of fluency, it is, for others, a measure of having become frustrated, given up, and moved to a new problem. The conclusion of this line of inquiry was to abandon time on task as a variable to be used in the intelligent system.

The questionnaire of the 1990 study concluded with a tabulation of percentages correct for task performance sub-areas (written, aural, keyboard) and a composite of performance on all tasks. To determine if raw scores had correlated with semester grades, I calculated average pretest scores for individuals in each semester grade group (A, B, C, D, F). Table 1 represents these percentages in a comparison of term grades with performance on the test prototype.

Table 1. Comparison of term grades with average pretest scores

	Average Pretest Scores, Fall 1990			
Semester Grades, Fall 1990	Written Subscore	Aural Subscore	Keyboard Subscore	Composite Score
Individuals with written grade of F	.67	.31	.48	.58
Individuals with aural grade of F	.70	.27	.56	.61
Individuals with overall grade of F	.68	.28	.53	.59
Individuals with written grade of D	.73	.56	.58	.67
Individuals with aural grade of D	.81	.55	.79	.77
Individuals with overall grade of D	.76	.41	.71	.70
Individuals with written grade of C	.78	.49	.78	.73
Individuals with aural grade of C	.80	.57	.69	.74
Individuals with overall grade of C	.80	.58	.77	.76
Individuals with written grade of B	.83	.60	.78	.79
Individuals with aural grade of B	.83	.68	.69	.78
Individuals with overall grade of B	.81	.63	.71	.76
Individuals with written grade of A	.83	.70	.89	.82
Individuals with aural grade of A	.81	.74	.75	.79
Individuals with overall grade of A	.78	.64	.82	.77

The table shows that students who failed the course achieved an aggregate raw score of 25% less than students who passed the course with an A and 20% less than students who passed the course with a D or higher. Thus, student performance in the test prototype did reflect grade distributions at the conclusion of the semester.

I also grouped pretest subscores into quintiles and compared them to average semester grades for individuals falling within each quintile. This comparison showed that students whose subscores were in the lowest fifth earned semester grades of D or F in theory. Students who scored in the highest fifth earned an average grade of B. Table 2 represents these figures (average semester grades of D or F appear in bold type).

The questionnaire's statistics for scale showed a mean of 68.56% with a standard deviation of 20.89%. Roughly two-thirds of the subjects scored between 48% and 88% in the composite raw score, roughly one-sixth scored below 48% and one-sixth scored higher than 88%. Inter-item correlation yielded a positive mean of .2972 with a minimum correlation (negatively discriminating item) of –.529, a maximum correlation (positively discriminating item) of 1.0, and a range of 1.529.

The most compelling statistic of the pretest, quantified in a variable called Cronbach's Alpha, is a measure of reliability and internal consistency more powerful than a split-half reliability. Cronbach's Alpha represents the percentage of time students who missed an item missed all harder items and students who correctly solved an item solved all easier items. Cronbach's Alpha for this study was .9727, indicating not only that the test was consistent with itself but that learning objectives in music theory are strongly hierarchical. For the purposes of the Ball State project, this last finding would prove to be extremely useful.

This research showed, moreover, that the hierarchy involves a correlation of aural and written preskills with each other. The data show particularly compelling evidence for the importance of aural skills. Students who received a low semester grade in written theory earned low subscores in both the written and aural portions of the test. Not only were both subscores less than for those earning A's, but the aural subscores were substantially lower. Whereas students who failed written theory achieved written subscores an average of

Table 2. Comparison of pretest distributions by quintiles with term averages for the fall semester, 1990

	Semester Grades, Fall, 1990			
Pretest (1st week of fall, 1990) Raw Scores Grouped in Quintiles	Written Average	Aural Average	Kbd Average	Overall Average
individuals scoring in the lowest 5th				
written tasks: 37%-65%	.67	.71	.86	.72
aural tasks: 0%-29%	.71	.65	.82	.68
keyboard tasks: 0%-40%	.66	.73	.85	.69
composite score: 30%-59%	.63	.68	.83	.68
individuals scoring in the second 5th				
written tasks: 67%-78%	.79	.79	.88	.76
aural tasks: 31%-44%	.76	.74	.88	.75
keyboard tasks: 41%-75%	.83	.82	.91	.82
composite score: 60%-72%	.83	.78	.91	.79
individuals scoring in the third 5th				
written tasks: 79%-83%	.84	.80	.92	.83
aural tasks: 50%-62%	.84	.81	.93	.83
keyboard tasks: 76%-85%	.80	.78	.88	.79
composite score: 73%-80%	.81	.83	.90	.83
individuals scoring in the fourth 5th				
written tasks: 84%-88%	.86	.79	.93	.82
aural tasks: 54%-79%	.87	.84	.93	.86
keyboard tasks: 86%-94%	.86	.78	.91	.86
composite score: 81%-85%	.88	.80	.94	.83
individuals scoring in the highest 5th				
written tasks: 89%-95%	.90	.86	.91	.87
aural tasks: 81%-100%	.90	.90	.95	.90
keyboard tasks: 95%-100%	.89	.83	.96	.85
composite score: 86%-96%	.91	.89	.95	.91

sixteen percentage points lower than their counterparts who earned A's, the same students' aural subscores were thirty-nine percentage points lower. Similarly, students who failed aural theory averaged eleven percentage points lower on the written subscore, but forty-seven points lower on the aural subscore, than those who earned A's. Students whose pretest written subscores were in the lowest fifth earned semester grades twenty-three points lower in written

theory and fifteen points lower in aural theory. Individuals whose aural subscores were in the lowest fifth earned semester grades nineteen points lower in written theory and twenty-five points lower in aural theory.

Accordingly, the most significant finding of the 1990 research was that aural preskills comprise a stronger basis for the assessment of preparedness for college theory than do written preskills. This conclusion corroborates a 50-year old study in which Taylor showed that aural skills correlate more highly with success in the music professions than do written skills.[6] If these conclusions are valid, testing ought well to focus upon aural skills, as ought the remediation (if not the theory sequence itself).

An Expert System

Advantages

It took the students of the pilot study nearly an hour to complete the computerized questionnaire. Because the content matrix of the pilot study proved to be hierarchical, it appeared that a branching instrument could dramatically accelerate the assessment process. If a student showing mastery at a high level was not compelled to advance through items lower in the hierarchy and vice versa, fully half of the test items could be eliminated. But to determine which half—the high or low end of the hierarchy—it was necessary to devise an intelligent system.

An intelligent system—sometimes called "expert system"—is software that interacts with the user in a nonlinear fashion based upon responses as they are integrated with, and interpreted by, data.[7] By the end of 1990, I thought that an expert systems test might do

---

[6]E. M. Taylor, "A Study in the Prognosis of Musical Talent," *Journal of Experimental Education* 10 ((1941): 1-28.

[7]Many computer-administered designs resemble paper-and-pencil tests mounted on a computer screen. Such was the case with Colman's, and, no doubt, similar instruments being engineered even now. While such adaptations do offer the advantage of automated administration and grading, they do little else to improve upon old technology, and use few of the unique capabilities of the new.

three things that a traditional test could not do: (1) generate problems algorithmically, (2) interact with the testee in real-time interpolation and analyses of data, and (3) progress in a nonlinear fashion to the most pertinent levels of the hierarchy, omitting the rest.

In addition, an intelligent system could accomplish at once two ordinarily separated diagnostic objectives: the ranking of students and the assessing of competence. One of the traditional imperatives of test design is to decide first if the purpose of the assessment is to rank students (norm-reference) or to identify competence (criterion-reference). This decision is necessary because each purpose requires a different type of problem. A norm-referenced test requires items of average difficulty and low (.25) to very high (1.0) discrimination. A criterion-referenced design intended to identify students at risk requires easy items and low discrimination (less than .25). Similarly, a criterion-referenced design intended to identify competence requires difficult items.

By contrast, an expert system could begin with norm-referenced items (do a preliminary ranking by sorting weak students from the strong), but conclude with criterion-referenced items (measure competence at both ends of the scale). By beginning with items of high discrimination the system would more quickly detect a trend toward mastery or nonmastery. After a trend had been established, substitution of criterion-referenced items would more efficiently and more accurately diagnose competence.

Three Learning Objectives and Competency Tests

Consider, for example, the following objectives and competency tasks related to theory assessment and how they might be ordered to yield information more efficiently and accurately.

Task One

The objective of Task One is for the student to demonstrate apprehension of the function of pitches in relation to a tonal center. Given a key signature and the first note of a seven-note diatonic melody (mostly steps and all whole notes), the student demonstrates competence by writing the remaining pitches correctly. The melody may be played an unlimited number of times.

This research showed that the percentage of students who could write any one pitch correctly, before having received instruction, averaged .83 for those who later passed the course but .33 for individuals who did not. For future reference, let us recall the probability that a master (Pm) would complete Task One correctly as .83, and the probability that a nonmaster (Pn) would complete the same task correctly as .33. As shown by the difference between Pm and Pn, Task One had high discrimination: D = .5. Accordingly, Task One would be appropriate at the beginning of a test where it might quickly sort weaker students from the stronger.

Task Two

The objective of Task Two is for the student to demonstrate apprehension of the aural structure of intervals of the second (major or minor). The computer generates a second, plays it starting on any pitch, and writes the first pitch in the student's favorite clef. The student is not required to name the interval but is instructed to represent the second pitch. Any diatonic or chromatic equivalent is recorded as a correct response. The student may play the interval an unlimited number of times. This study showed that Pm for this type of problem was .27 for students who later passed theory and .08 for students who did not. The discrimination variable was .19.

Because Task Two does not have high discrimination it is not appropriate at the beginning of the test (where we wish to distinguish between weaker and stronger students). But because Task Two is difficult, we shall save it for students whom we suspect belong to the latter group.

Task Three

The objective of the third task is for the student to name the symbols of notation. The computer names a symbol and displays four. The student is instructed to choose the display that represents the named item. This study showed that the probability of correctly selecting the symbol was .96 for students who later passed theory and .90 for students who did not: Pm = .96,

Pn = .9, and D = .06. Like Task Two, this task is unsuitable for the beginning of the test because it does not have high discrimination. But, because this task is easy, we shall save it for students whom we suspect are insufficiently prepared.

Discrimination and Item Difficulty as Parameters Controlling the Efficiency and Effectiveness of Assessment

The goal of a computer-adaptive test is to facilitate a more efficient and more accurate identification of students at risk. Efficiency is a measure of the time needed to accomplish that goal and accuracy is a measure of the validity of the assessment. The specified ratio of correct to incorrect responses may be adjusted depending upon how efficient and accurate one wishes the assessment to be. If one desires a faster but less accurate assessment, one might accept a lower ratio of successes to failures.

For example: if our subject correctly completes Task One one time we should begin to suspect mastery. However, because one-third of the nonmasters could also complete this particular task correctly, we should not be confident of our assessment until the student has replicated correct responses to a specified ratio, depending on how accurate we want the assessment to be. We must therefore continue to press the testee with different versions of Task One. Each new success, assuming there are no failures, will make us more confident of our assessment, but a mix of successes and failures might lead us to conclude that we dare make no conclusions on the basis of Task One by itself.

As mentioned before, specific success to failure ratios are not without inverse effects upon the efficiency and accuracy of the assessment. Generally, the faster the assessment, the less accurate, and vice versa. Equally significant, whatever success to failure ratio is specified, it must be calibrated with item discrimination. Items with higher discrimination require a lower proportion of successes; items with lower discrimination require a higher proportion. The following four paragraphs elaborate upon this principle.

With Task One, the number of successes that we might require depends upon the discrimination potential of the task, in this case already very high. If the discrimination for Task One had been still

higher—say, 1.0 (all of the masters, but none of the nonmasters, had solved it correctly)—then we might have concluded that our assessment was correct after one success. If the discrimination had been lower—say, .1 (masters solved the task correctly 10% more often than did nonmasters)—then our subject would need to accumulate roughly nine successes for every failure before we could be confident of an assessment of mastery.

Task Two is more difficult than Task One, but it has less potential to discriminate. If our student correctly completes Task Two one time we may be more confident that he is a master than if he solved Task One one time. Inversely, if our student misses Task Two one time, we must be less confident that he is a nonmaster than if he missed Task One one time. The second task is, after all, much more difficult. Because eight percent of nonmasters could also complete this second task correctly, we cannot presume mastery until we determine what kind of success to failure ratio our student can produce when given a series of problems like Task Two. Because Task Two has lower discrimination than Task One we will need a higher ratio of successes to failures to conclude mastery.

Task Three is easier than Task One, much easier than Task Two, and has the least potential to discriminate. This does not mean, however, that Task Three is of no value—quite to the contrary. Because the purpose of the test is to identify students at risk, Task Three is exactly the type we require. Of course, we will present tasks like number three only to students who we suspect are at risk, and to make this preliminary assessment we need tasks like number one.

Correctly completing Task Three one time says practically nothing about competence, as nearly all nonmasters could do the same. But, if our student misses Task Three one time, we should strongly suspect nonmastery. By missing Task Three we can be more sure that our subject is a nonmaster than if he or she had missed the more difficult first or second tasks. Because the discrimination index for Task Three is low, we must accumulate more than the usual number of failures to conclude nonmastery.

Five Variables

In an expert system we can envision, therefore, five variables governing the weight that we should ascribe to any given task. The first three of these are statistical functions and the last two are accumulated during the taking of the test itself: (1) Pm is the percentage of those in the pilot study who passed the course and who were able to solve the task correctly before having had instruction, (2) Pn is the percentage of those who failed the course who were able to solve the same task correctly before having had instruction, (3) D, the difference between Pm and Pn, is the discrimination variable, (4) S is the number of real-time successes at any given task that the test may present, and (5) F is the number of failures.

Besides these variables, the efficiency and accuracy of an expert system might be controlled by predetermined ratios—successes to failures that serve as error factors, a priori. The first of these variables, Alpha, shall represent the percentage of time we are willing to tolerate the misidentification of masters as nonmasters. The second, Beta, shall represent the percentage of time we are willing to tolerate the misidentification of nonmasters as masters. High Alpha and Beta result in a faster (but less accurate) assessment.

Probability, Decision, and Branching Structures and the Sequential Probabilities Ratio (S.P.R.)

If the expert system envisioned in 1990 were to exist, it required a mathematical rubric that could keep track of its variables, interpolating their significance in real time, and revising probability estimates with each success or failure. Such a rubric was found in Bayesian mathematics, a branch of statistics devoted to the continuous revision of probabilities until reliable estimates have been established. Developed originally for military applications, the equation used in this study was declassified and published after World War II. Since that time it has been used for quality control in manufacturing. Abraham Wald's equation has only recently come to the attention of educators thanks to the work of Indiana University's Theodore Frick. [8]

---

[8]Theodore Frick, "A Comparison of Three Decision Models for Adapting the Length of Computer-Based Mastery Tests," *Journal of Educational Computing Research* 6 (1990): 483.

Because the formula is abbreviated S.P.R.—for Sequential Probabilities Ratio—Frick has coined the term, "ExSPRT system," for "Sequential Probabilities Ratio Test." The equation follows:

Figure 1. Equation for computing sequential probabilities ratios

$$PR = \frac{Pm^{S}(1-Pm)^{f}}{Pn^{S}(1-Pn)^{f}}$$

Following each completion of a task, the computer recalculates PR and compares it to three logical structures that determine if the subject is a master or nonmaster of the task in question. If **PR ≥ (1-ß)/a** then the student has demonstrated mastery of the objective. If **PR ≤ ß/(1- a)** then the student has demonstrated nonmastery of the objective. If **ß/(1-a) < PR < (1-ß)/a** then it is not possible to determine yet whether the student has or has not mastered the objective; generate another problem.

Ready or Not?

Table of Specifications

With the discovery of the S.P.R. Equation it was a relatively simple task to design the expert system envisioned at the conclusion of the pilot study. I began by selecting 22 learning objectives for which problems could be generated at random on a computer. These included a mix from difficult to easy and from high to low discrimination. Each was empirically quantifiable for its statistical significance in the longitudinal study of 1990-'91. Content specifications for the ExSPRT system are reproduced in Table 3 (arranged from highest to lowest discrimination). Notice that seven of the 22 objectives (asterisked items) assess aural acuity and that these seven comprise the objectives with the highest discrimination and/or difficulty.

Table 3. Ready or Not: Table of specifications

Twenty-two learning objectives (highest to lowest discrimination)	*Pm*	*Pn*	*D*
1 writes short diatonic melody*	.83	.33	.50
2 chooses aural option that correctly plays polyphonic notation*	.81	.47	.34
3 writes enharmonic equivalents	.85	.53	.32
4 writes pitches in treble, alto, and bass clefs	.85	.54	.31
5 plays pitches in treble and bass clefs on keyboard	.79	.48	.31
6 hears tonal fragment and writes last pitch in given key*	.69	.39	.30
7 knows the number of sharps and flats in key signatures	.50	.20	.30
8 identifies the root of a triad	.56	.26	.30
9 chooses aural option that correctly plays monophonic notation*	.74	.47	.27
10 completes a measure with missing durations	.74	.50	.24
11 identifies longest/shortest of four difficult rhythms	.80	.60	.20
12 identifies starting pitch of a song*	.50	.30	.20
13 writes M/m melodic seconds*	.27	.08	.19
14 names pitches in treble, alto, and bass clefs	.93	.75	.18
15 hears triad and ids type (M,m,+,d) congruent with notation*	.22	.05	.17
16 writes ♯s/♭s of key signature in correct order and position	.85	.68	.17
17 identifies nonequivalent rhythms	.86	.69	.17
18 identifies longest/shortest of four easy rhythms	.87	.72	.15
19 understands the function of sharps and flats	.82	.69	.13
20 identifies notes and rests that are correctly paired	.94	.86	.08
21 names basic symbols of notation	.96	.90	.06
22 identifies the number (but not the quality) of intervals	.93	.88	.05

As currently configured, Ready or Not (RON) begins by selecting one of the five most discriminating objectives and generating random problems within that objective. As each student demonstrates mastery or nonmastery of the objective RON advances to one of the top five remaining objectives. If RON is unable to determine mastery or nonmastery after an instructor-prescribed number of tasks within each objective, it advances to a new objective.[9]

When RON detects a trend toward mastery or nonmastery (as defined by a collective probability ratio for all tasks in the range of >.66 or <.33 respectively), it reorders the objectives remaining in its

[9]To prevent the assessment from becoming too long, the instructor may prescribe a maximum number of tasks in any one category (ordinarily a dozen). If RON is unable to make an inference after this maximum, it indicates, for the record, that it was unable to inform itself as to the student's aptitude on the basis of that objective.

matrix appropriately. If RON suspects, for example, that a student is a master it will begin to select items from the more difficult objectives remaining in its content matrix. Because of the high Cronbach's Alpha of the pilot study RON can assume, with confidence, that after a student demonstrates mastery of one or two difficult objectives the assessment can be discontinued. RON assumes, in such a case, that the student does not require remediation and should therefore be advised to register for the first course in the sequence.

Results

The School of Music at Ball State University has been using Ready or Not, with satisfactory results, for the past three years. Students are normally assessed during the senior year of high school at one of several audition dates. During the visits they also play their instruments before a jury of the faculty, write an essay, and are interviewed. A small stream of individuals is continuously assessed throughout the year on a walk-in basis. The computer administers the assessments, interprets the results, and advises its subjects in one sitting. The advice ranges from indicating that the subject may register for Music Theory 111, may register but should prepare during the summer by studying Harder and Steinke's *Basic Materials in Music Theory*, or should not begin the sequence without having taken the remedial course. Students in this last category are not restricted from enrolling in the regular sequence if they so choose. Approximately one-sixth of the applicants are advised to take the remedial course (roughly the failure rate before implementation of RON). Most students follow RON's advice. As a consequence, enrollment in the remedial course has more than doubled while the failure rate in the sequence has dropped.

During the first year of its use RON surprised even its author with the rapidity with which it completed its assessments (often in less than ten minutes and after having assessed only two or three objectives). Because I wished the students to have a more substantial engagement with the material, in the second year RON was constrained to withhold prognosis until after it had assessed a minimum of five objectives.

The S.P.R. structure dictates that students at the extremities of the curve—those who are either extremely deficient or well pre-

pared—will be identified more quickly than students who are average. In the case of the former, even after the minimum coverage of five objectives, RON often detects a need for remediation within fifteen minutes or less. But, in the case of average performance, some students continue through the entire table of specifications with RON being unable to formulate a prognosis. Consequently RON has been programmed to allow the student to work no longer than thirty minutes (or some other instructor-specified range). If it is unable to prognosticate the need for remediation within the half hour, RON advises the student to enroll in the first semester of the sequence.[10]

Case Studies

Robert[11]

RON assessed Robert in January of 1992, generating the report reproduced in Figure 2. In the fall of 1992 Robert enrolled in Music Theory 111 and passed it with an A. He recently completed the fourth semester of the sequence, having maintained an A average throughout.

Notice that RON tested Robert in but five areas. RON began by randomly selecting three of the more discriminating objectives (those from the top of Table 3). Robert was first asked to write certain pitches in specified clefs (objective 4 from Table 3). This particular objective had a discrimination factor of .31. This means that of the students in the pilot study who had mastered the objective before having had instruction, 31% more passed the course than failed it.

RON continued by notating pitches in various clefs and prompting Robert to play them on the keyboard—again a high discrimination factor of .31 (objective 5 from Table 3). For the third segment RON queried Robert about the number of sharps and flats in major

[10]The instructor may define this time limit and several others, including the controlling error factors Alpha and Beta.

[11]Individuals identified in this study are real students who were assessed by Ready or Not in the spring of 1992. Their names have been changed to protect confidentiality.

Figure 2. RON Report for Robert, January 25, 1992

Learning Objectives MASTERED:
1. writes specified pitches in treble, alto, and bass clefs
2. plays specified pitches in treble and bass clefs on keyboard
3. knows the number of sharps and flats in key signatures
4. writes M/m melodic seconds
5. matches sounding triad (Mm+o) with written triad

Learning Objectives NOT MASTERED: (none)

Learning Objectives INCONCLUSIVE: (none)

Prognosis:
Robert appears to have MASTERED the preskills that are necessary for success in the first year of music theory. The way this session was configured, there is a 16% chance that Robert might actually be UNPREPARED. Elapsed time for this session: 11 minutes.

Enrollment Management Ranking: **5**
1 = DEFINITELY needs remediation
2 = probably needs remediation
3 = may not need remediation, but advised to brush up anyway
4 = prepared for MusTh 111
5 = EXCEPTIONALLY well prepared for MusTh 111

and minor key signatures—a discrimination factor of .30 (objective 7 from Table 3).

At this point RON had already prognosticated that Robert was ready for Music Theory 111. Because it had been constrained to cover a minimum of five content areas before delivering its prognosis, RON proceeded with the assessment, but with a different strategy. It selected its final two objectives from those that could most quickly confirm its prognosis of mastery—difficult items. For the fourth segment Robert was required to write the intervals that he

heard: major and minor melodic seconds (objective 13 from Table 3). Only 27% of those in the pilot study who passed theory were able to do this, and Robert was able to demonstrate mastery during his assessment.

Finally, RON displayed a series of triads and asked Robert to match the sounds (M, m, +, or o) with what he saw (objective 15 from Table 3). Only 22% of those in the pilot study who passed theory were able to do this, as was Robert. Clearly Robert was ready for the theory sequence, a conclusion substantiated by his subsequent performance in four semesters of study.

Notice that RON made its assessment of Robert—an unusually well-prepared student—in eleven minutes. Notice, too, that RON has informed us of the error factor, Alpha, that it used in its assessment. There was a 16% probability that Robert was actually not prepared. Because RON's purpose was to identify students at risk, the instructor was willing to tolerate a high Alpha in order that the test might not become too long.

Melissa

RON assessed student Melissa in March of 1992, generating the report reproduced in Figure 3. At the conclusion of the session RON advised Melissa that she should enroll in the fundamentals course before attempting the sequence. That fall Melissa enrolled in the remedial course and failed it. She has since dropped out of the major.

We note that Melissa had not mastered any of the objectives that RON presented. As in Robert's assessment, RON first engaged Melissa with tasks of high discrimination (objectives 1 and 3 from Table 3 with respective discrimination factors of .5 and .32).

When RON began to suspect that Melissa was not adequately prepared, it sought to confirm its prognosis by pressing her with easier tasks that she, likewise, had not mastered. Melissa could not negotiate objective 17, for example ("identifies nonequivalent rhythms" from Table 3). This is a skill that more than two-thirds of the individuals in the pilot study who failed the first semester of theory could do. Neither was Melissa able to name pitches in various clefs (objective 14), a skill that three-fourths of the unprepared students in the pilot study could do before having received instruc-

Figure 3. RON Report for Melissa, March 14, 1992

Learning Objectives MASTERED: (none)

Learning Objectives NOT MASTERED:
1. writes short diatonic melody
2. writes enharmonic equivalents
3. identifies nonequivalent rhythms
4. names pitches in treble, alto, and bass clefs

Learning Objectives INCONCLUSIVE:
1. understands the function of sharps and flats

Prognosis:
Melissa appears NOT to have mastered the preskills that are necessary for success in the first year of music theory. The way this session was configured, there is a 7% chance that Melissa might actually be PREPARED. Elapsed time for this session: 13 minutes.

Enrollment Management Ranking: 1
1 = DEFINITELY needs remediation
2 = probably needs remediation
3 = may not need remediation, but advised to brush up anyway
4 = prepared for MusTh 111
5 = EXCEPTIONALLY well prepared for MusTh 111

tion. It is clear that, when she was assessed in the spring of 1992, Melissa was not prepared to begin the theory sequence. She followed RON's advice and enrolled in the remedial course, which she failed.[12]

RON completed its assessment of Melissa in thirteen minutes. Notice that the error factor, in Melissa's assessment, was 7% as op-

[12]None of the teachers of our 25 theory classes and labs or remedial courses except myself has access to RON assessments of students coming into their courses. I do not teach the remedial course.

posed to Robert's error factor of 16%. There was a 7% probability that Melissa might actually have been prepared. Because the purpose of the assessment is to identify students at risk, Ball State uses a lower Beta than Alpha. This means that RON's estimates of nonmastery are more accurate than its estimates of mastery—exactly what we want.

Thirteen Students Identified as Needing Remediation in 1992

Including Melissa, RON had, by the middle of March, 1992, identified thirteen prospective students as either definitely or probably needing remediation. Three students ignored RON's advice and registered for the first course in the sequence without having been remediated. Two of these withdrew before the eighth week, and one failed the course. (Although I have no evidence that the two students who dropped were failing the course, students typically drop for this reason.)

Four of the thirteen did not matriculate in the major. They may have become discouraged by RON's prognosis, changed their majors, or they may not have enrolled in the University at all. The remaining six students who were advised to enroll in the remedial course followed the advice. Of these six, two students failed (Melissa was one), one withdrew, one received a D, and two passed the course with a B. Of the two B students, one continued into the regular sequence and passed the first semester with a grade of C; the other did not continue in the sequence (at least not at Ball State).

Where Do We Go From Here?

In the current instrument, Pm and Pn are static, representing a synchronous view of performance in the five institutions that participated in the 1990 study. It is postulated, however, that both variables are diachronically and environmentally sensitive. The proportion of masters and nonmasters who will pass or fail a course having mastered prescribed objectives before instruction surely depends upon institutional admissions policies, grading scales, and curricula, if not individual teaching styles. A test, therefore, that might be accurate at one site might be less or more accurate at another site, or at the same site years from now. I am continuing to

revise Ready or Not to devise a test that learns—a test that remembers the performance of its subjects, revises its own estimates and rewrites its own code to provide a dynamic assessment of institutional patterns in time.

As the Ball State School of Music continues to use Ready or Not into a fourth year, the effectiveness of the system is being studied for its accuracy as an aid to advising. I envision Ready or Not as a tool that could be used not only by departments of theory at the college or university levels, but also by private instructors and high school career counselors.

If the S.P.R. approach to theory assessment is a viable strategy for the identification of students at risk, further research is necessary. Many more preskills could be used for which I currently have no data (Pm and Pn) with which to incorporate them into Ready or Not. It is my hope that by presenting the results of this research others will be encouraged to conduct investigations leading to an expanded and more accurate assessment.