

Predictive Validity of the “Ready or Not” System for the Assessment of Students Needing Remediation in Music Theory

Timothy A. Smith

A perennial problem in the teaching of music theory is how to identify students who are insufficiently prepared for the harmony/ear-training sequence. If it is a logistical problem—how to convene, grade, and advise such a throng before registration, it is also a problem of assessment—what skills and knowledge should one test, how exactly should a student proficient with the major key signatures (but not minor) be advised, and how poorly must one perform to be steered onto the remedial route?

To solve the first problem, some instructors have devised computerized versions of paper-and-pencil tests. While such assessments have the advantage of walk-in administration, online grading and advising, they do not ordinarily employ the more powerful computations of so-called “expert” systems: non-linear and branching pathways, algorithmic computation of probabilities, content shaped by in-progress assessment, random generation of problems, databases, empirically informed cutoffs or norm-referenced advisement.

Given the hierarchical organization of basic music theory, the computer is the perfect tool for tailoring test content, in real time, to the perceived strengths and weaknesses of individual students. Having interpolated reliable estimations from failed attempts at easier problems, for example, intelligent systems are easily engineered to skip more difficult problems where students have little or no probability of success.¹

¹The objective is a faster, but no less reliable, assessment. In research leading to the creation of “Ready or Not,” it was shown that a high percentage of students who missed an item missed all harder items (and students who solved an item solved all easier items). That measure, as quantified by Cronbach’s Alpha, a statistical indication of internal consistency more powerful than a split-half reliability, was .9727.

In 1994 the *Journal of Music Theory Pedagogy* included a report of a project undertaken by this author to design and implement such a system.² That article described research preceding the creation of the instrument as well as ExSPRT control structures which enabled it, normally, to identify target populations in under half an hour.³ The system has since come to be known by the acronym RON (Ready or Not). The RON system was programmed to select skills to be tested from the objectives listed in Table 1; examples of stimuli for presentation and measurement of selected objectives are presented in Appendix II.

RON begins by selecting at random an objective with high discrimination; can the student, to cite one example, notate pitches that are enharmonically equivalent. RON first provides a brief explanation of "enharmonic," then writes a random pitch (in the student's favorite clef) and instructs the student to write its equivalent. The visual stimulus for this type of problem is represented in Figure 1.

RON was designed to present items in response to the test taker's performance on preceding items. Following each attempt, RON would employ algorithms to calculate the probability that the student has mastered this particular objective.⁴ When RON is able to

²"Timothy A. Smith, "An ExSPRT Systems Approach to the Assessment of Students Needing Remediation in Music Theory," *Journal of Music Theory Pedagogy* 8 (1994): 179-200.

³The "SPRT" of "ExSPRT" stands for "Sequential Probabilities Ratio Test." The test uses Bayesian mathematics continuously to revise estimates until a reliable prediction can be made, at which time the test is concluded. The mathematical equation at the heart of this system may be found in the next footnote.

⁴The following equation computes a Probabilities Ratio (PR). Because this is done after each problem, the algorithm may be understood to compute Sequential Probabilities Ratios (or SPR, see fn. 3 *supra*). The equation follows in which: Pm = the percentage of students passing the course (in the 1990 study) and who could do this type of problem before having had instruction, Pn = the percentage of students NOT passing the course who could do this type of problem before having had instruction, S = successful answers by the current testee, F = unsuccessful answers by the current testee. (Pm and Pn for "writes enharmonic equivalents" were 85% and 53% respectively.)

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

Table 1: Possible Test Items in RON

completes a measure with missing durations
 identifies last pitch of a phrase in C Major
 identifies correctly paired notes/rests
 identifies longest/shortest of four easy rhythms
 identifies non-equivalent rhythms
 identifies starting pitch of a song
 identifies the root of a triad
 knows number of sharps and flats in key signatures
 matches sounding triad (M m + d) with written
 names basic symbols of notation
 names pitches in treble alto and bass clefs
 plays pitches of treble and bass clefs on keyboard
 understands the function of sharps and flats
 writes enharmonic equivalents
 hears and writes M/m melodic seconds
 writes pitches in treble alto and bass clefs
 writes sharps/flats of key sig. (correct order and position)
 writes short diatonic melody (dictated)

make a reliable prediction, it leaves the objective for another.⁵ If a student masters problems of high discrimination, RON switches to a more difficult cluster, where the prognosis of mastery is more quickly confirmed. Conversely, if RON's preliminary assessment indicates that a student may be poorly prepared, it moves to easier

Following each problem, the value of PR is compared with three evaluative statements in which: α = error tolerance for misidentifying masters as non-masters, and β = error tolerance for misidentifying non-masters as masters.

If $PR \geq (1-\beta)/\alpha$ then the student has demonstrated mastery of the objective.

If $PR \leq \beta/(1-\alpha)$ then the student has demonstrated non-mastery of the objective.

If $PR \beta/(1-\alpha) < PR < (1-\beta)/\alpha$ it is not possible to make a reliable prediction, then give the student another problem of the same type.

⁵In instances where failed attempts to answer the question balance successful attempts, RON terminates the line of questioning and ranks the student as "inconclusive" on that particular objective. Inconclusive assessments on individual items do not count for, or against, the student in the overall assessment.

Figure 1: Stimulus for High Discrimination Item
(Writes Enharmonic Equivalents)



**Write this pitch
ENHARMONICALLY.**

items where a lack of preparedness can be more quickly established.⁶ As one can see from the foregoing description, RON is non-linear and unrepeatable in precisely the same format.

While the 1994 article did establish the internal consistency of the test in a measure known as Cronbach's Alpha (fn. 1 *supra*), it did not, other than to present a handful of case studies, make any assumptions about predictive validity, sufficient data being at that time unavailable. The current article represents analyses of data accumulated since then with a view to establishing the predictive validity of RON for identifying students at risk.

While this article reports on important information for teachers of music theory and individuals involved in the creation and administration of placement tests, it is possible, without sufficient context, that the current research may be difficult to interpret without reference to the companion article (*JMTP* 8, 1994). The introduction has, hopefully, redacted sufficient information to render the conclusions of this investigation understandable. However, readers with a particular interest in the mechanics and pedagogical issues of non-linear tests may wish to study the earlier report (see fn. 2).

The target population for this investigation was comprised of 400 Ball State University students commencing a major in music

⁶Demonstrated mastery of difficult items (or non-mastery of easier items) allows the test to be concluded without moving through irrelevant items in the test inventory. Please refer to fn. 1, *supra*.

from the fall of 1992 through 1995. Data were at hand for 281 students assessed by RON as high school seniors and who subsequently matriculated as music majors, completing a music theory course for grade. No data existed for the more than one hundred students who enrolled without advisement.⁷

Four out of five individuals in the data group completed the first course of the music theory sequence while the other fifth completed the remedial course.⁸ Inasmuch as the purpose of this study was to assess the predictive validity of RON as it pertained to performance in the music theory sequence, with the exception of those whose remedial attempts resulted in failure, students enrolled in the remedial course were excluded from consideration.⁹

Because half of the students advised into remediation effectively self removed from the sample by having heeded RON's advice the data were inherently biased.¹⁰ Inasmuch as these individuals chose

⁷Because the data pool was not selected at random, it was not possible to infer properties of the larger population from known values. The sample does, however, properly adduce attributes to a sub-group of high school seniors with the means and inclination to participate in on-site assessments prior to matriculation. Conclusions of this report with respect to predictive validity may therefore be considered appropriate in instances where RON has been utilized prior to matriculation, preferably as high school seniors.

⁸Although two fifths of sampled students were advised into the remedial course, half of them opted to disregard RON's notice, attempting the first course in the sequence outright. More than two thirds of this latter group failed.

⁹Throughout this report failure of the course is defined as having earned a grade of D or F. Although D was a passing grade at Ball State, students who earned D's in the theory sequence were required to repeat the course. It seemed reasonable to assume that students who failed remediation would have also failed the first semester of the sequence. Accordingly, the sample retains data on students who engaged themselves in unsuccessful attempts at remediation.

¹⁰Whereas the statistical ideal would have required denial of student access to the results of their RON assessments, enrollment of students in the sequence regardless of preparedness, would have been unconscionable in view of what was, from case studies, already known about the utility of the instrument. The inclusion, in the data pool, of failed attempts at remediation represented an effort to compensate for bias by reconstituting a reasonable approximation of the ideal population.

not to leave themselves in a position where lack of preparation might have been established empirically, inaccurate estimations of this study would be expected to have erred on the conservative side.

As it was first constructed, RON assessed students either as prepared, unprepared, or inconclusive. For the purposes of this study, the data justified a six-point ranking (Table 2 below) comprising the predictive variable. The outcome variable was represented by the student's grade at the end of the semester.

The bivariate data of this study were readily adapted to two-way contingencies like that of Table 3. While these data indicate

Table 2: Predictive Variable (quantified by RON rankings of 1-6)		
RON score		
Predictive Variable	Qualification for Rank	Prognosis
1	identified by RON as prepared, having mastered at least 85% of objectives	exceptionally well prepared for Music Theory
2	identified by RON as prepared, having mastered 70 - 84% of objectives	prepared for Music Theory
3	identified by RON as prepared, having mastered at least 70% of objectives	may not need remediation, may be prepared
4	ran out of time (or objectives) without RON having made a prognosis	may need remediation, may not be prepared
5	identified by RON as unprepared, having mastered at least 15% of objectives	probably needs remediation
6	identified by RON as unprepared, having mastered less than 15% of objectives	definitely needs remediation

that RON assessments correlate with term grades, the direction of correlation cannot be taken for granted.¹¹

The selection of an appropriate measure to interpret these data required a determination of the purpose of the measure—in this case, to establish the validity of RON's predictions. Available measures depended, in part, upon whether the data were ranked in a continuum of infinite range (scalar) or partitioned into discrete units (categorical). Both the semester grade and RON's prognosis are, of course, categorical rankings. As one can see from Table 2, RON

Table 3: RON Scores by Grade

Recommendation	Grade					
	Frequency					
	Total Percent					
	Row Percent					
Recommendation	Column Percent					
	A	B	C	D	F	Total
Advised into Mus Th Sequence (RON score 1,2,3)	28	34	42	22	14	140
	11.24	13.65	16.87	8.84	5.62	56.22
	20.00	24.29	30.00	15.71	10.00	
	93.33	73.91	67.74	39.29	25.45	
Advised into Remediation (RON score 4,5,6)	2	12	20	34	41	109
	0.80	4.82	8.03	13.65	16.47	43.78
	1.83	11.01	18.35	31.19	37.61	
	6.67	26.09	32.26	60.71	74.55	
Total	30	46	62	56	55	249
	12.05	18.47	24.90	22.49	22.09	100.00

¹¹While any symmetrical value could have been used to establish correlation, the direction of the correlation could not, in such instances, be assumed. Most comparisons, for example, produced Chi-Square values of less than .0005 indicating statistically significant correlation between variables. Such correlation could not indicate, however, which one, or both, of the variables was dependent. Thus, Chi-Square could not properly be used to distinguish between RON assessments as estimators of term grades or the reverse. Table 4 presents data that does indeed establish the direction of correlation.

scores are the product of asymmetrical cutoffs applied *a priori* to prognoses of mastery or non-mastery.¹²

Three measures commonly used to establish statistical relationships are Goodman-Kruskal Gamma (γ), Kendall's Tau-b (τ_b), and Somers' D. The problem driving each of these measures is: for any randomly-selected pair of students, what is the probability that one student's RON score and subsequent term grade will be concordant, discordant, or tied with the other's.¹³ This analysis employed Somers' D as the appropriate measure because it was capable of making finer distinctions between variables even in instances where one or both were tied.¹⁴

¹²Such rankings do not allow one to quantify how much more of a master a student ranked 1 is over a student ranked 2. Similarly, whereas a semester grade of F could have represented a range of from 0-69%, all other grade units were comprised of 15% increments. While it is possible to assert that an F is lower than a B, it is impossible to determine how much lower. While the six categories of the predictive variable were not scalar, RON calculated them by means of probability ratios of from zero to one hundred, with an infinite range of decimals between. Product of a Bayesian equation at the heart of the ExSPRT system (see fn. 4 *supra*), a continuous scale was reconstructable, given extant data, and permitted a comparative analysis predicated upon scalar, rather than ordinal, formations. It was hypothesized that validity might be improved by means of ratio variables admitting quantitative comparison. Such a comparison showed, unexpectedly, that the categorical format was, in this instance, more reliable.

¹³Goodman-Kruskal Gamma represents a conditional probability, the condition being a lack of ties. Kendall's Tau-b and Somers' D differ from Goodman-Kruskal Gamma, primarily, in how they recognize ties. Kendall's Tau-b accounts for ties as a value equal to the geometric mean of all types of tied observations. Kendall's Tau-b is, therefore, an unconditional probability lending itself specifically to the analysis of ordinal bivariate data where ties are likely to occur. The disadvantage of Kendall's Tau-b is that it is incapable of making distinctions between the different types of ties. Of the three measures under consideration, only Somers' D distinguishes between ties having ramifications as to the independence of variables with ties accruing to either variable, or both. It does this by expressing itself in two directions. In terms of asymmetric, or two-way, contingencies, these values are commonly phrased as row-predicts-column, versus column-predicts-row.

¹⁴For example, Joe's RON score may be the same as Jane's but their term grades may be different. In this instance the two observations are

Somers' D for the data represented in Table 3 is .519 (see Table 4 below). This value is indicative of a moderate relationship between RON's categorical rankings and subsequent grades earned in the first semester of harmony/ear-training. Perhaps of more significance, this figure shows that the RON score is in proper orientation to the term grade. In other words, RON functioned as it was designed: to rank order students as per a value congruent with eventual performance in the music theory course.¹⁵ But how reliable was this rank order, and can it be used to predict performance?

Measurements of Somers' D may range from -1 to 1. These values indicate, respectively, that every pair of observations is either discordant or concordant. A measure of zero indicates lack of correlation, with neither variable manifesting statistical evidence of contingency upon the other. Positive values are indicative, therefore,

Table 4: Predictive Validity of RON's Asymmetric Categories				
	cut-off point used to predict			
	...adequate preparation	...need for remediation	Somer's D	Asymptotic Standard Error
Asymmetric Categories (1-6 RON rankings)	1-3	4-6	.519	.059
	1-4	5-6	.517	.072

concordant if the RON score is viewed as the independent variable but discordant if the term grade is viewed as independent. By contrast, Joe's and Jane's term grades may be the same, but their RON scores may be different. Such a tie would indicate concordance with respect to the outcome, but discordance with respect to the prediction. Finally, Joe's and Jane's RON scores and term grades may be identical, indicating a degree of concordance without implying the independence of either variable.

¹⁵This can be illustrated by reversing the direction and using the final term grade to "predict" the RON score. With such an orientation, Somers' D drops to .323. Please note that what is being described here is not a cause-and-effect relationship. The RON score did not cause the term grade. What is being described, instead, is a statistical measure showing a moderate relationship between variables allowing one to conclude that a student's performance in the sequence may mirror the RON score.

of correlation in which the predictive measure is in proper orientation to the outcome. Such a value is more reliable the further removed it is from zero.¹⁶ Somers' D of .519 would indicate, therefore, a borderline acceptable degree of predictive validity.¹⁷

How much confidence should one place in such a figure? Some discussion is in order. As a predictive measure, RON scores accounted for 26.9 percent (.519²) of the variance in the music theory grades with 5.9 percent prediction error. It is important to emphasize that the positive linear correlation (Somers' D .519) implies simply that low RON scores can be associated, to a degree, with poor performance in a music theory class. In other words, the "prediction" is not that a student with a low RON score will fail music theory. It is, rather, that in any population of students who have taken RON, the rank order that it produces will correspond, to a certain extent, with the eventual term grade distributions in the music theory class.

But how well does RON identify students at risk? If RON ranks at-risk students at the top, then students at the lower ranks would also be at risk. Conversely, if RON places well-prepared students at the bottom, then students in the higher ranks should outperform them. So, while a correlation coefficient of .519 may warrant some degree of confidence in RON's rankings, it does not indicate where the cutoff should be and whether RON merely happened upon that value. At what point should a student be required to take the remedial course?

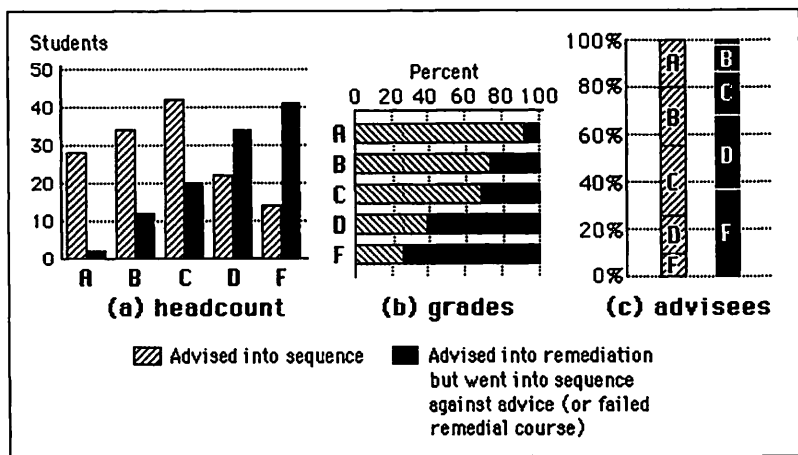
¹⁶So, if Somers' D for RON had been +1, RON assessments could have been used to predict that every student with a higher score than his fellows would have earned a higher term grade in the sequence than his fellows. Conversely, if Somers' D for RON had been -1, RON assessments could have been used to predict that every student with a LOWER score than his fellows would have earned a HIGHER term grade in the sequence than his fellows (i.e. reliable as a predictor, but in the wrong direction). If Somers' D for RON had been zero, it could not have been used to predict anything.

¹⁷A correlation coefficient of less than .50 is generally considered to be invalid for predictive purposes. The Somers' D value of .519 falls within the range of acceptable. In statistical analyses of this type, a prediction error of 5% or less would have been considered acceptable. While RON's prediction error of 5.9% is slightly high, it does approach the accepted level and should decline as the test continues to be developed.

In answer to the question of cutoffs, the reader is referred to Table 5. The first view (5a) represents a comparative headcount of students whom RON advised into the remedial course versus those advised into the music theory sequence. The second view (5b) represents percentiles of remedial vs. non-remedial students within each grade bracket. The third view (5c) compares grade distributions of those advised into remediation vs. those advised into the music theory sequence. Note that more than two thirds of the failing grades were earned by students whom RON assessed as needing remediation (Table 5b). Similarly, more than two thirds of those advised into remediation either failed the first semester of the sequence or the remedial course itself (Table 5c).¹⁸

As noted earlier, RON's prognosis was represented in categorical rankings based upon asymmetrical cutoffs (see fn. 12 *supra*). To determine if a continuous scale might establish a more reliable cut-off, probability ratios for each student were converted to a 1-10 scale,

Table 5: Three Comparisons of Outcomes Between Students Advised into the Sequence Versus Students Advised into Remediation



¹⁸These percentages are, of course, exclusive of individuals who followed RON's advice to remediate and who subsequently passed the preparatory course.

contributing not only four more categories, but also a truly symmetrical scale capable of rendering quantitative distinctions.¹⁹ Using this new estimator, Somers' D values were calculated and compared with the old, yielding the symmetric categories of Table 6.

Table 6: Comparison of the Validity of Asymmetric vs. Symmetric Categories				
	cut-off point used to predict			
	... adequate preparation	... need for remediation	Somer's D	Asymptotic Standard Error
Asymmetric Categories (1-6 RON rankings)	1-3	4-6	.519	.059
	1-4	5-6	.517	.072
Symmetric Categories (1-10 Continuous Scale)	1-9	10	.496	.063
	1-8	9-10	.482	.061
	1-7	8-10	.459	.062
	1-6	7-10	.434	.062

Table 6 shows, surprisingly, that while the decision to delineate the predictive variable (RON score) in asymmetric categories was intuitive, it happened upon a linear function that better predicts outcomes both in terms of Somers' D and minimization of Asymptotic Standard Error (hereafter referred to as ASE). Whether the remedial line is drawn at 4 or 5, assessments are more reliable using RON's asymmetrical categories than the reconstructed symmetrical scale.

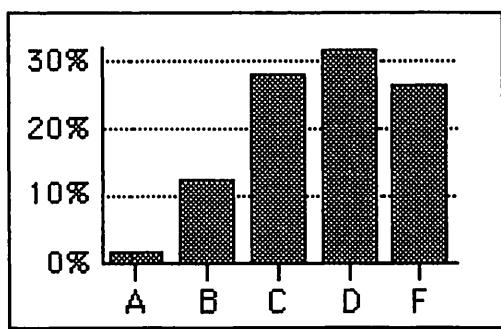
In addition to showing that asymmetrical categories are more reliable, Table 6 suggests that the predictive validity of RON would not be unduly compromised if students with RON rankings of 4 are advised either direction: to remediate or proceed with the first se-

¹⁹While RON's asymmetrical assessment was not expressed in continuous variables, it was possible to reverse-engineer Sequential Probabilities Ratios (see fn. 4 *supra*) to reconstruct the values RON used to determine if students were prepared. These values comprised a continuous spectrum of from zero to one hundred with an infinite range of decimals between.

mester of music theory.²⁰ Had students with RON rankings of 4 been advised into the first semester, Somers' D would have been .517. But with the same group advised to remediate, Somers' D was .519. The minimal difference (.002) in validity is negated by possible ASE of .072 and .059, respectively. Inasmuch as the validity of the instrument would not have been compromised, Table 7 shows that it is clearly in the interest of these students to remediate. Fifty-eight percent of their peers who proceeded to take the first semester of harmony/ear-training, without such treatment, ultimately failed. Table 7 represents grade distributions for this group.

RON was engineered to render assessments after having evaluated at least five objectives. Accordingly, it should be understood that the Somers' D value of .519 discussed to this point represents RON's predictive validity as a composite of five or more learning

Table 7: Grade Distributions. Students with Inconclusive Assessments Overall (RON Rank 4)



²⁰Recall that RON reserved the rank of 4 for individuals for whom it could make no reliable estimate of preparedness within the allotted time (see Table 2). In other words, the rank of 4 technically represents no prognosis at all—at least not from RON. But this is not to say that these students could not be ranked. Because this research established that students assessed as inconclusive overall performed at a level between those whom RON had placed into what are now identified as rankings 3 and 5, the category of 4 was created to represent them.

objectives. In other words, discussions of validity, till now, have focussed upon RON as an instrument, and not upon the validity of its individual learning objectives.²¹ We examine now the criterion validity of each objective. These shall be presented roughly in order from the least, to the most, reliable. Somers' D values for each objective as predictor of the term grade (and vice versa) may be found in Appendix I.

RON's assessment of the ability to identify the last pitch of a heard phrase in C Major produced the only negative Somers' D value of the study. When assessments of this skill alone were used to predict passes and failures, Somers' D was -.164, suggesting that a student might be even more likely to pass the course if RON assessed him as a non-master of this objective! The high ASE (.235) indicates that this prognosis is suspect. Because omission of the objective effects a slight decrease in Somers' D for the instrument, it would appear that it does indeed contribute somewhat to the overall validity.²²

Two of the objectives appear, at first glance, neither to have added to, nor subtracted from, the assessment. Of individuals asked to identify correctly paired notes and rests, for example, one was as-

²¹Although experimentation with a continuous scale did not produce a more reliable predictor than RON's asymmetrical categories, the symmetrical approach was instructive in that it enabled calculation of predictive reliabilities (as expressed in Somers' D) for individual learning objectives (*vis a vis* the RON instrument overall). These values are detailed in the Appendix I which shows that individual objectives as predictors of the term grades were, in every case, more accurate than the reverse. The scalar predictor allowed, too, for comparisons of hypothesized configurations in which one or more learning objectives had been removed: e.g. to what extent would RON have been more, or less, reliable if Objective No. 1 had not been part of the system.

²²It is possible that one defect may be the manner in which the objective is contextualized. When told that they will hear a phrase of music in the key of C Major, students are instructed to write the last pitch. The purpose of the problem is to assess the extent to which the testee perceives tonal centricity and scale degree function. No explanation of these functions is given, it being assumed that individuals who possess the skill will understand, intuitively, what they are being asked to show, and show that they can do it by writing the last pitch they hear. It is possible that with additional explanation this line of assessment might yield better results.

essed as not having mastered the objective, while all others were assessed as inconclusive.²³ Similarly, every person asked to name the basic symbols of notation was assessed as inconclusive.

The 1990 phase of this research showed that both of the above objectives were achievable by a high percentage of persons who eventually failed the first course in the music theory sequence.²⁴ It was thought, therefore, that persons tested as non-masters of these particular objectives would be poor candidates for music theory. While inconclusive assessments of individual objectives did not count against students, RON assessed 77 percent of those who tested inconclusive in their ability to match notes and rests as unprepared on the basis of deficiencies in other areas (63 percent of these students ultimately failing the course). Similarly, RON ranked 75 percent of those who tested inconclusive in their ability to name the basic symbols of notation as unprepared because of lack in other pre-skills (61 percent of them ultimately failing the course).

That sampled students to whom the above two objectives were directed showed themselves, overall, to be unprepared, indicates that RON functioned as expected. This holds true in spite of the fact that most were assessed as inconclusive with respect to these two objectives. It bears repeating that the sample omitted individuals whom RON advised to remediate and who completed that remediation successfully. Of this latter group, nearly one third who were tested on these two objectives were assessed as not having mastered them. Thus, if students advised to remediate (and who heeded RON's advice by enrolling in the remedial course and passing it) were actually unprepared at the time of assessment, then these two objectives were efficacious.²⁵

²³The student assessed as not having mastered this objective earned a grade of D in the sequence.

²⁴Of those who failed, 86% could pair notes and rests (compared to 94% of those who passed), and 90% could name the basic symbols of notation (compared to 96% of those who passed).

²⁵There was no way, from this study, to have established this empirically. Recall that half of the students advised to take the remedial course followed that advice, and half did not (fn. 8 *supra*). Only the half that ignored the advice (and those who followed it but failed the remedial course) were included in this study. It is this group that was assessed as

RON's assessment of a student's ability to write pitches in treble, alto, and bass clefs predicted student performance with low Somers' D of .083. The ASE of .083, indicates that this assessment may be suspect.²⁶ If it is reliable, low congruity shows a limited contribution to the overall assessment. Omitting this objective from the analysis, Somers' D for RON assessments increases slightly indicating that the removal of this objective might increase the validity of the system.²⁷

Of the forty-eight individuals required to match sounding triads (M, m, +, d) with written triads, all but two tested as masters. The two exceptions tested as inconclusive. The 1990 phase of research found that 22 percent of students who passed, but only 5 percent of those who failed, the first semester of music theory could do this before having had instruction. Unlike easy objectives, which were presented to students whom RON suspected as being unprepared, this difficult objective was presented to students whom RON suspected as being prepared. The purpose of the objective was to confirm the initial prognosis quickly, after which the assessment could be concluded.

A traditional evaluation of the effectiveness of individual test items might have concluded that, whereas the ability to match sounding triads with written did not distinguish between those who were prepared and those who were not (all but two students testing as masters), it should be eliminated. It turns out, however, that 71

inconclusive on these two objectives. Of the group that did follow RON's advice, took the remedial course and passed it, nearly one third were assessed as not having mastered these objectives. It was not possible to establish that this latter group was truly unprepared (as RON had said it was) without them having failed something. Instead, they passed the very remedial course that RON had advised them to take.

²⁶Although it is unusual that a correlation coefficient would be the same as its ASE, in this instance the values were indeed found, after repeated calculation, to be identical!

²⁷Before deleting the objective, further research should be conducted to verify the original estimates of mastery and non-mastery with respect to writing pitches in various clefs. Whereas the 1990 research produced different estimates for each clef, RON was designed to merge these data, selecting clefs at random, and averaging estimates with each problem. If the original estimates were correct, the method used for merging data may be inaccurate.

percent of these students evaluated as having mastered the objective were assessed, on the basis of other testing, as prepared for the music theory sequence overall. Seventy-seven percent of this group passed the course. Once again, the idiosyncratic design of the ExSPRT system was such that this type of problem was given primarily to students who actually were masters. The effectiveness of such a strategy is revealed, ultimately, in the fact that overall predictive validity of the system, as shown in Somers' D, drops when this objective is omitted.

RON's assessment of a student's ability to identify triad roots produced a low Somers' D value of .111. The ASE (.116), being higher than the estimate, allows a possibility that the estimate may be off. When this skill was used to predict actual grade distributions, Somers' D increased to .184 with an ASE of .125. When this objective is omitted, predictive validity of the assessment decreases by two percent. This suggests that the objective contributes to the validity of the system, overall, and that it should be retained in spite of its performance as an individual indicator.²⁸

Of the objectives identified in the 1990 phase of research, the ability to write, from dictation, a short diatonic melody promised to distinguish most quickly between prepared and unprepared students. The current research yielded Somers' D of .249 for this objective as indicator of actual grade distributions (ASE.126). Were RON to have omitted this objective Somers' D would have dropped by 1.5 percent. Whereas the objective does contribute to the validity of the system overall, it appears to be less dramatic than originally anticipated.

RON made no inconclusive assessments of any of the 211 individuals it required to notate from dictation a short diatonic melody. Of all learning objectives, this one alone identified every student as either proficient or deficient, indicating that the original estimate of spread between the two was excessive. Wide estimates typically

²⁸Unlike the previous objective, where too little set-up information may have been problematic, the problem here may have been too much information. Students who did not know what a triad root was were given the opportunity to take a tutorial before attempting the assessment. It is possible that omitting this tutorial may give a more reliable estimate.

cause the ExSPRT system prematurely to conclude assessments, and therefore not as reliably as it should.²⁹ Another possibility is that RON's diatonic melodies, which were all stepwise, were too easy compared to melodies of the original research (which contained some leaps of a third). RON has subsequently been revised to play melodies including thirds.

The ability to demonstrate understanding of the function of sharps and flats predicted grade distributions to an accuracy of Somers' D .446, with an ASE of .115. It should be noted that, in addition to testing the student's understanding of how accidentals function, this item also evaluated the student's understanding of whole-steps and half-steps. Because the setup information appeared to confuse some students, this research was expected to show that the question was unreliable, or at least required modification. Surprisingly, the data show that this objective is the fourth most reliable predictor of performance.³⁰

With Somers' D of .522 and an ASE of .173, the ability of students to identify the longest or shortest of four easy rhythms was the third most reliable objective. After being presented with four rhythms, three of which spanned the same number of beats, students were asked to indicate which of the four was longer (or shorter, as the case may be).

The second most reliable objective was the ability of students to identify the starting pitch of a song. Here the student was told to write, in the key of C, the starting pitch of a folk song. Students were allowed to opt out of tunes they did not know. The correlation coefficient for this objective was .586, with an ASE of .132.

²⁹The 1990 research showed that 83% of students who passed the first semester of theory could write short diatonic melodies, but that only 33% of those who failed could do it.

³⁰The testing of this objective required explanation that appeared to puzzle students and sometimes required verbal clarification. The setup begins with the computer writing a pitch in the student's favorite clef. This pitch may be unaltered, double-flatted, flatted, sharped, or double-sharped. To the left of the staff a grid represents these signs, and the student is instructed as follows: e.g. "Choose the symbol that, when substituted for the given symbol (if any), would produce a pitch one HALF step HIGHER" (or LOWER, depending upon the question).

The most reliable objective was the ability to hear and to write melodic seconds.³¹ Here the computer displayed the first pitch in the student's favorite clef then played a second pitch a major or minor second removed. Students were required to notate the second pitch (enharmonic equivalents were counted as correct). Students could play the interval as many times as they wished. With low ASE of .065 and .078 respectively, Somers' D for this objective was .824 as predictor of term grades and .706 as predictor of pass/fail.

Using its asymmetrical rankings, the validity of RON overall was established at Somers' D .519. This value was met or exceeded by only three objectives: student writes M/m melodic seconds (.824), student identifies the starting pitch of a song (.586), and student identifies the longest/shortest of four easy rhythms (.522). The mean Somers' D value for individual objectives was .370. While .370 (average for individual objectives) to .519 (validity overall) is not a sta-

³¹It might puzzle some that the three objectives with the highest validity are the ones with the fewest students tested. The objection might be raised, for example, that if writing M/m melodic seconds is so much more reliable than anything else, might one conclude that this is all that is really needed in such a test? The reason so few students were given these particular questions is because of the unique strategy employed by the ExSPRT system. These problems were difficult ones: used only to confirm a high level of competence in instances where prior queries had determined that the student was probably a master. For example, the 1990 pilot study showed that only 27% of those who passed the first semester of theory could write M/m melodic seconds before having had instruction (and a mere 8% of those who failed the first semester could do this). So, what would happen if this were the only objective of the test? Answer: a very low percentage of students would demonstrate mastery of the objective, and one could be quite confident that they were prepared for music theory. But that is not the object of the assessment. The purpose of RON was to identify students NOT prepared for theory. And the 1990 research showed that 73% of even the PREPARED students had not mastered this particular objective before having had instruction. Accordingly, non-mastery of this objective is of very little use in determining which students are unprepared for the sequence. Thus the expert system asks this question only of persons it suspects, from prior questioning, to be masters—hence the low numbers.

tistically rigorous comparison, it does support what some might consider to be obvious—that the cumulative effect of testing across many objectives increases the validity of the system.³²

Throughout this discussion it has been noted that RON occasionally found it impossible to assess a student as having mastered, or not mastered, one or more objectives. In such instances the student was evaluated as “inconclusive” as pertaining to the objective in question (see fn. 4 & 5 *supra*). While RON currently disregards inconclusives—as if the student had never attempted that type of problem—this research enabled the asking of the question: what, if anything, was measured when RON assessed a student’s competence with respect to individual learning objectives, as inconclusive?

It had been assumed, in the creation of the ExSPRT system, that inconclusives were actually masters, or non-masters, but that RON had collected insufficient information to make such a distinction. This research suggests, however, that inconclusive assessments of individual objectives are indicative of a level of competence somewhere between mastery and non-mastery. Mean scores of students testing inconclusive at any given objective bear this out. The system may be improved by factoring inconclusive assessments of individual objectives as somewhere between mastery and non-mastery.

Two final observations are in order. First, the data that makes RON “work,” as well as data for this study showing that it worked, were obtained from institutions that attract students of average preparation for university-level music theory.³³ With the exception of the handful of volunteers from Indiana University (in the 1990

³²The comparison is not statistically rigorous because certain objectives were called upon with less frequency than others. For example: 211 individuals were required to write short diatonic melodies (Somers’ D of .249), while only 35 were required to write M/m melodic seconds (Somers’ D of .824).

³³Data from the 1990 pilot study were obtained from 141 volunteer freshmen at Ball State University, Indiana University, Biola University, Taylor University, and the University of Oregon. Data for this study in predictive validity were obtained from students at Ball State University. Data continue to be gathered and analyzed from students (since 1995) at Northern Arizona University.

pilot study), the average RON testee might be described as below that of, shall we say, most nationally ranked conservatories and schools of music. In the decade over which RON has evolved, it is likely that this particular clientele has been "hard wired" into the system. In the last eight years the author has directly administered RON, or supervised its administration, to more than 1,300 students at Ball State University and Northern Arizona University. RON has rank ordered these students within standard deviation limits of the normal curve.

It is certain that if the same instrument were used at institutions where the average student is more (or less) prepared, the curve-linear profile of the instrument would be quite different. It is conceivable that this alteration could render the utility of RON, in some instances, ineffective—all students are ranked 1 or 2, for example. This is not to invalidate, however, the testing approach. With some adjustments to its specifications, objectives, and cutoffs, RON could very well be adapted to serve institutions of different size, mission, and student preparedness. Research at unlike institutions would be required, however, for any adjustments to be well-informed.

Second, the notion that some students, who were assessed as unprepared, actually passed the course—two with an A and twelve with a B—might be a little worrisome for those who are not used to these types of assessments. The converse might be easier to comprehend. Of the 140 students whom RON assessed as prepared to begin the music theory sequence, thirty-six did not pass (22 earning D's, and 14 earning F's).

In coming to terms with anomalies, it is important to remind ourselves that students fail courses for reasons other than lack of pre-skills, which explains why no assessment can be expected to be completely reliable. Lack of discipline, or money, or reasons of emotional or physical health, learning disabilities—factors entirely ignored by RON—would explain most of the inconsistencies. Too, RON had no mechanism for detecting that extra ounce of intelligence, that dogged determination, abnormal curiosity, or unusual self-discipline, that might have compensated in instances where it had identified students with insufficient pre-skills (which might explain, though not entirely excuse, those two A's and twelve B's of the preceding paragraph).

Ultimately, we must ask, are the seventy-five individuals whom RON identified as deficient (and who subsequently failed their first music theory course) worth the misidentification of fourteen who proved RON “wrong” by earning A’s and B’s in the first semester of music theory? The answer to that question is most likely “yes.” In view of the fact that many students advised to remediate followed that advice, and passed their first music theory course, the proportion of students whom RON spared “theoretical trauma” (not to mention waste of time and money) outweighs its relatively infrequent misdiagnoses.

This research is significant to music theory pedagogy in five ways. First, it establishes the validity of the intelligent design concept for alerting advisors to students who may not function well in a music theory course unless they receive special attention. Since RON began as an experiment in a radically different concept in assessment, this outcome is very good. With continued work, the prediction error should be reduced and the predictive validity of this instrument, and others like it, improved. Second, this research contributes to our understanding of pre-skills that are more (or less) indicative of success in the music theory sequence. Third, it adds to the growing body of evidence that the rudiments of music theory are hierarchical and that “magic bullet” assessments—where success at higher levels justifies skipping the lower—may be efficient indicators of competence. Fourth, it helps us to know where cut-offs should be made in the advising process. Specifically, we know from this research that students who mix successes with failures while attempting any given objective (especially where that “style” of answering persists over three or four objectives) are probably not prepared.³⁴ Fifth, while this particular assessment was designed to identify students as unprepared for beginning music theory and ear-training, perhaps of greater importance is the efficacy of computer-adaptive assessments over a wide array of learning objectives throughout the music theory curriculum. Could, for example, an ExSPRT system identify students who might be allowed to skip

³⁴This type of response characterized the performance of the “out-of-timers”—those whom RON assessed as inconclusive overall and who were, for the purposes of this research, ranked 4.

lower division courses? This research suggests that, where the purpose is to achieve an accurate assessment in as short a time as possible, and where learning objectives are highly hierarchical (as in music theory) it is possible to design computer-administered assessments that are capable of predicting outcomes with a high degree of validity.³⁵

³⁵The author thanks Dr. Roy T. St. Laurent, statistics instructor and consultant at Northern Arizona University, for his help in designing and executing this research. Also of invaluable assistance were Dr. Graydon W. Bell (NAU), and Dr. Patricia Sink (the University of North Carolina, Greensboro), who provided most constructive criticisms pertaining to the final report.

APPENDIX I

Somers' D values Per learning objective
as predictor of term grade and vice versa.

(ASE in parenthesis)

Objective	Somers' D				Students Tested
	Objective Predicts Term Grade	Objective Predicts Term Grade	Term Grade Predicts Objective	Term Grade Predicts Objective	
	Pass /Fail	All Gdes	Pass /Fail	All Gdes	
completes a measure with missing durations	.319 (.100)	.332 (.112)	.297 (.095)	.201 (.069)	95
identifies last pitch of a phrase in C Major	-.164 (.235)	.181 (.235)	-.073 (.109)	.052 (.070)	33
identifies correctly paired notes/rests			See note #1		
identifies longest/shortest of four easy rhythms	.362 (.215)	.522 (.173)	.253 (.163)	.217 (.092)	29
identifies non-equivalent rhythms	.282 (.145)	.377 (.157)	.248 (.130)	.216 (.093)	51
identifies starting pitch of a song	.507 (.126)	.586 (.132)	.507 (.126)	.363 (.084)	41
identifies the root of a triad	.111 (.116)	.184 (.125)	.080 (.084)	.085 (.059)	103
knows number of sharps and flats in key signatures	.281 (.096)	.302 (.105)	.229 (.080)	.156 (.056)	127
matches sounding triad (M m + d) with written			See note #2		
names basic symbols of notation			See note #3		
names pitches in treble alto and bass clefs	.189 (.123)	.160 (.138)	.189 (.123)	.105 (.090)	64
plays pitches of treble and bass clefs on keyboard	.201 (.075)	.304 (.081)	.196 (.073)	.183 (.049)	180
understands the function of sharps and flats	.360 (.111)	.446 (.115)	.357 (.110)	.289 (.075)	71
writes enharmonic equivalents	.272 (.086)	.358 (.103)	.173 (.058)	.142 (.044)	196
hears and writes M/m melodic seconds	.706 (.078)	.824 (.065)	.091 (.087)	.063 (.060)	35
writes pitches in treble alto and bass clefs	.083 (.083)	.151 (.093)	.078 (.079)	.088 (.054)	154
writes sharps/flats of key sig. (correct order and pos.)	.302 (.105)	.380 (.113)	.300 (.104)	.240 (.071)	83
writes short diatonic melody (dictated)	.197 (.109)	.249 (.126)	.074 (.043)	.059 (.031)	211

(notes)

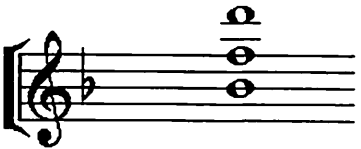
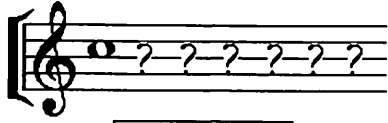
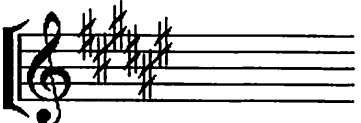
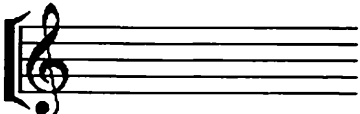
1) It was not possible to compute statistical values for this objective as most individuals fell within the same cell. Of the 39 individuals tested all but one were inconclusive. The remaining person tested as not having mastered it.

2) Not possible to compute statistical values: of the 48 individuals tested, all but two tested as masters. The remaining two tested as inconclusive.

3) Not possible to compute statistical values: all 44 individuals tested in this objective were inconclusive.

APPENDIX II

Stimuli for Selected Objectives

 Click on the ROOT of the triad.	 Play DONE Write the melody.								
 add Flat Done add Sharp Write the signature for e minor.	<table border="1" data-bbox="627 572 929 720"> <tbody> <tr> <td></td> <td></td> </tr> <tr> <td></td> <td></td> </tr> </tbody> </table> Click on the item (above) that completes the measure (below).								
 In the key of F-sharp Major Write the first pitch of <i>Joy to the World</i>	$\frac{12}{16}$ 								
<table border="1" data-bbox="201 1041 504 1206"> <tbody> <tr> <td></td> <td></td> </tr> <tr> <td></td> <td></td> </tr> </tbody> </table> Click where notes and rests are correctly paired.					Click on the line or space that is E <table border="1" data-bbox="616 1093 924 1223"> <tbody> <tr> <td></td> <td></td> </tr> <tr> <td></td> <td></td> </tr> </tbody> </table> Click on the BASS CLEF sign				