DOCUMENT RESUME

ED 312 314 TM 014 121

AUTHOR Vispoel, Walter P.; Twing, Jon S.
TITLE A Comparison of the Efficiency, Reliability and Validity of Adaptive and Conventional Listening Tests.
PUB DATE Mar 89
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (San Francisco, CA, March 27-31, 1989).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Adaptive Testing; *College Students; Comparative Testing; Higher Education; High Schools; *High School Students; Individual Testing; *Listening Comprehension Tests; *Music; Talent Identification; Test Format; Test Reliability; *Test Validity
IDENTIFIERS *Music Ability; Seashore Measures of Musical Talents; Tonal Memory

ABSTRACT

The measurement precision, efficiency, and validity of an adaptive test and four conventional listening tests designed to assess musical ability were compared. The conventional tests were the Seashore Tonal Memory Test and three tests (peaked, rectangular, and maximum discrimination) constructed from items in the 278-item adaptive test pool. The results were based on data from 468 high school and college students. The 30-item adaptive test provided comparable or, in the vast majority of cases, superior measurement precision to the conventional tests at all ability levels. Measurement precision comparable to the conventional tests was achieved by the adaptive test using 34% to 69% fewer items. Although differences tended to be small, in most cases adaptive test validities exceeded those of the conventional tests. The findings suggest that adaptive testing procedures, which, prior to this study had been limited to written items, can provide significant improvements in the measurement of listening skills. Five tables and three graphs provide study information. (Author/SLD)

A Comparison of the Efficiency, Reliability and Validity of Adaptive and Conventional Listening Tests.

Walter P. Vispoel

Jon S. Twing

The University of Iowa

Abstract

The purpose of this study was to compare the measurement precision, efficiency, and validity of an adaptive test and four conventional listening tests designed to assess musical ability. The conventional tests were the Seashore Tonal Memory Test and three tests (peaked, rectangular and maximum discrimination) constructed from items in the 278-item adaptive test pool. The results were based on data from 468 high school and college students. The 30-item adaptive test provided comparable or, in the vast majority of cases, superior measurement precision to the conventional tests at all ability levels. Measurement precision comparable to the conventional tests was achieved by the adaptive test using 34% to 69% fewer items. Although differences tended to be small, in most cases adaptive test validities exceeded those of the conventional tests. The findings suggest that adaptive testing procedures, which prior to this study had been limited to written items, can provide significant improvements in the measurement of listening skills.

Introduction

When the assumptions of Item Response Theory hold and a large pool of appropriate items exists, in theory computerized adaptive tests offer potential solutions to three fundamental problems that often plague conventional fixed-item tests--namely, poor efficiency, reliability and validity. Compared to conventional tests, adaptive tests should be: a.) more efficient because examinees respond only to items matched to their ability level, b.) more reliable because examinees receive a greater proportion of items at the proper difficulty level, and c.) more valid because increases in test reliability should lead to increases in test validity. The purpose of this study was to compare the efficiency, reliability and validity of conventional and adaptive tests in an area in which no previous adaptive testing research has been done--that of listening skills.

Background and Theoretical Framework

The theoretical advantages of adaptive tests have been demonstrated in studies by McBride and Martin (1983), Urry (1977), and Weiss (1982), who compared adaptive and conventional tests of similar content. Taken as a whole, these studies indicate that adaptive testing procedures can reduce significantly the number of required items without reducing test score reliability and validity. Across these studies, adaptive tests required an average of only 5 to 9 items to reach reliability estimates of .80 and resulted in a need for 50% to

80% fewer items to achieve reliabilities comparable to conventional tests. Furthermore, in the McBride and Martin (1983) study, concurrent validities for adaptive test scores were significantly higher than those for conventional tests up to 19 items in length.

It should be noted that the vast majority of studies on adaptive testing (including those cited) have been restricted to written tests of verbal (e.g. vocabulary) and numerical ability. Recent state-mandated public school testing programs (e.g., Public Act 84-126, State of Illinois, 1985), however, call for the assessment of a broader range of skills, including listening and speaking. It could be argued, in fact, that the development of adaptive tests in the area of listening is more important than in other areas. The reason for this is that fatigue becomes a problem more quickly in listening tests. Since listening items are presented only once (or at most twice), any lapse in attention may result in an incorrect item response. This problem is less serious in written tests because examinees are able to reread a given item as many times as they desire before responding. To minimize the effects of fatigue or lapses in concentration, then, efficient but reliable and valid tests are particularly important in the listening domain.

In this study, listening skills were assessed using tests of "tonal memory". Tonal memory has been shown to be an important component of musical ability (e.g. Whellams, 1971). This study has five specific objectives:

1. To compare the measurement precision of a 30-item adaptive test to several fixed 30-item conventional tests and three 30-item tests formed from the adaptive test item pool.

2. To compare the relative efficiency of the adaptive test to each of the conventional tests as a function of ability level.

3. To determine the average number of adaptive and conventional test items required to yield various levels of reliability (i.e. .80, .85, .90 and .95).

4. To determine the average reliability level of adaptive and conventional tests at various fixed test lengths (5, 10, 15, 20, 25 items etc.).

5. To compare validity coefficients for the adaptive and conventional tests at various fixed test lengths using several criterion measures, including standardized tests of musical ability and indices of musical experience and training.

Methods and Data Source

The present study uses the data from which an operational computerized adaptive test of tonal memory was developed (See Vispoel, 1987). In the previous study, items for the adaptive test were constructed to parallel those of commercially available tests, notably, the Seashore et al., (1960), the Wing (1968), and the Bentley (1966) tonal memory tests. A typical test item requires an examinee to compare two versions of a short melody. On the second playing, either the same melody is repeated or one note is altered. Examinees indicate the number of the note that has changed or indicate that the two playings of the melody are the same.

The final item pool for the adaptive test contained 278 such items selected to provide high-fidelity measurement across all ability

levels. Item parameter estimates, based on data from 468 high school and college students, were obtained using LOGIST V (Wingersky, et al, 1982) under the modified three-parameter logistic model (i.e., the c parameter estimates were fixed). These items were tested for unidimensionality and model fit in the previous investigation. In addition to responding to the items in the pool, the majority of examinees completed the Seashore Tonal Memory Test, the Drake Musical Memory Test and an investigator-designed questionnaire assessing musical background and training. These measures served as criterion variables in the investigation of test score validities and are described in more detail in Table 1.

In the present study, the data from the Vispoel (1987) study were used to simulate computerized adaptive tests of various lengths. In these simulations, items were selected using a maximum information item selection procedure (see Hulin, Drasgow and Parsons, 1983 pp. 221-222). In this procedure, the first item on each test is the most discriminating item available at average difficulty. That item is scored immediately and ability is estimated. Next, the item that provides maximum information at the estimated ability level is administered and ability is reestimated. The third item is chosen to provide maximum information given the new ability estimate. This process is continued until the test reaches the desired length (5 items, 20 items, etc.).

In addition to the adaptive tests, three types of fixed-item conventional tests (flat, peaked and maximum discrimination) were

developed from items within the adaptive test pool. In a fashion similar to McBride and Martin (1983), test lengths identical to the simulated computerized adaptive tests were generated by identifying subsets of items that for the flat test were most discriminating across a wide range of abilities, for the peaked test were highly discriminating in the middle of the ability range only, and for the maximum discrimination test were the most discriminating items in the pool independent of difficulty level. More specifically, the flat conventional tests were generated by sorting the 278 items from the adaptive test pool into the following five difficulty categories: 1) $b \leq -1.5$, 2) $-1.5 < b \leq -.5$, 3) $-.5 < b \leq +.5$, 4) $+.5 < b \leq +1.5$, 5) $+1.5 \leq b$. The 5-item flat test contained the most discriminating item from each respective difficulty category, the 10-item flat test contained these five items plus the second most discriminating item from each difficulty category and so on. The peaked conventional tests were generated by alternately administering the most discriminating item available with a "b" value less than zero followed by the most discriminating item available with a "b" value greater than zero. The conventional maximum discrimination tests were generated by administering the items with the highest available "a" parameter values. The average a, b, and c parameter values for items within the 278 item pool were 1.18, 0.07, and 0.14 respectively.

The scores derived in the present study were based on a combination of simulated and actual data. Ability estimates for the conventional and adaptive tests were based on simulated responses.

These responses were obtained as follows: Each examinee's "true ability" level was defined as the ability est.. .e derived from his or her responses to 95 items from the original 278 item pool. Based on the item parameters and the examinee's "true ability", the probability of the examinee answering a given item correctly was computed directly from the equation for the three parameter logistic model (see Hambleton and Swaminathan, 1985, pp. 49). This probability value then was compared to a random number selected from a uniform (0,1) distribution, to determine whether the item was scored as correct or incorrect. For example, if the probability of getting an item correct is .32, then the item would be scored correct if the random number falls between 0 and .32 and incorrect otherwise. This procedure is repeated until an adaptive or conventional test reaches a termination criterion. Depending upon the analysis, the present tests were terminated after either a specified number of items were administered, or an examinee's ability estimate reached a predetermined level of reliability. To assess the accuracy of this simulation procedure, ability estimates from the actual 95 items that each examinee responded to and ability estimates from simulated responses to these same items were correlated. The obtained correlation of approximately .98 suggests that the two procedures yield nearly identical results. In contrast to the simulated scores on the investigator-designed adaptive and conventional tests, criterion variable scores (i.e. standardized musical ability test score and musical training experience indices were based on the actual responses of examinees.

Results

Measurement Precision: The measurement precision of the 30-item adaptive test was compared to the measurement precision of four conventional 30-item tests. Thirty-item tests were employed for three reasons. First, the Seashore Tonal Memory Test has 30 items. Second, 30-item tests have been used in previous investigations comparing conventional and adaptive tests (e.g., Crichton, 1981; Weiss, 1982; McBride and Martin, 1983). Finally, 30-item tests should provide a conservative basis for comparing the measurement quality of adaptive and conventional tests because research has shown (Urry, 1977; Crichton, 1981; Weiss, 1982; McBride and Martin, 1983) that the differences in measurement precision between adaptive and conventional tests decreases as test length increases. Figure 1 shows test information curves for the 30-item adaptive test and four 30-item conventional tests, the investigator-produced flat, peaked, and maximum-discrimination tests and the Seashore Tonal Memory Test. Information values for the conventional tests were computed directly from the item parameters and the ability values (See Lord, 1980 p. 74, equation 5-6). Items from the Seashore Tonal Memory test were calibrated to the adaptive test item pool scale using a weighted least squares equating procedure developed by Haebara (1980). In a manner similar to Crichton (1981), information values for the adaptive test were based on simulated responses of 100 examinees at each of 17 levels of theta ranging from -3.2 to +3.2 in intervals of .4. Adaptive test

information values were computed as the sum of the item information values for the items administered to each simulee, averaged across the 100 simulees at each theta level.

The information curves for the peaked and flat tests from Figure 1 follow an expected pattern given how the tests were constructed. As evident from Figure 1, test information for the peaked test is highest at theta levels near zero and drops off rapidly as theta moves away from zero. The flat test provides less information than the peaked test at middle ability levels (i.e., theta's between -7.0 and 7.0) but more information at extreme levels. The information provided by the maximum discrimination test falls somewhere between the peaked and flat test--better than the peaked test at the extremes, poorer at the middle; better than the flat test at the middle and poorer at the extremes. The Seashore test information is highest at a theta value of about -1.4, and drops to zero at theta levels greater than +.4. The information curves for the conventional tests illustrate the difficulty in constructing reasonable length fixed-item tests that provide high measurement precision across all ability levels. Test developers usually are forced to compromise measurement quality by choosing to measure certain ability levels with high precision and sacrificing precision at the other levels.

Such compromises are not necessary with an adaptive test, as reflected by the differences between the information curves for the present adaptive and conventional tests. It is evident from Figure 1 that over most of the ability range, the adaptive test provides higher

measurement precision than any of the conventional tests. These differences are even more dramatically illustrated by the relative efficiency plots in Figure 2. Relative efficiency for a given conventional test was computed by dividing the adaptive test's information by the conventional test's information. A relative efficiency of 1.0 indicates that the adaptive and conventional test provide equal measurement precision at a given theta level. A relative efficiency of 2.0 indicates that the adaptive test is twice as informative as the conventional test at the given theta level, a relative efficiency of 3.0 indicates that the adaptive test is three times as informative, and so on. Another way of interpreting these relative efficiencies of 2 and 3 is to say the conventional test would require respectively twice and three times as many items to reach comparable degrees of measurement precision. As evident from Figure 2, comparable measurement precision to the adaptive test was obtained by the Seashore test only at ability levels of -2.7 to -1.8, by the peaked test only at ability levels near zero, and by the flat test only at ability levels between 2.4 and 2.7. If one assumes that ability scores are normally distributed, these results would indicate that the adaptive test provides better measurement precision than the Seashore, peaked and the flat tests over approximately 97%, 84% and 99% of their respective ability distributions. The maximum discrimination test failed to provide measurement precision comparable to the adaptive test at any ability level.

The information curve for the Seashore test should be of special interest to individuals familiar with the test. As can be seen from Figure 1, the test provides its best measurement precision at very low ability levels and clearly is not well matched to the ability levels of the present sample. The low information values at moderate to high theta levels suggest that the test reaches a ceiling at medium ability levels. Not surprisingly, the test has a mean number correct score of 25.2 with a standard deviation 4.62, and 16.2% of the examinees obtained a perfect score of 30. In addition, 66% of the present examinees were within one standard deviation of a perfect score. It is clear from Figure 1 that no such ceiling effects are present for the adaptive test. Elimination of potential floor and ceiling effects is another distinct advantage of adaptive testing procedures. It also should be noted that the maximum level of measurement precision of the Seashore is no greater than that of the flat test, and the Seashore's bandwidth is considerably narrower.

Efficiency: Table 2 shows the average number of test items required to yield reliabilities of .80, .85, .90 and .95 for the adaptive test and the peaked, flat and maximum discrimination conventional tests. Reliability was based on a formula suggested by Urry (1977) and others (e.g., Green et al, 1984) in which reliability is defined as one minus the reciprocal of test information, i.e.:

$$r^2_{(\theta,\hat{\theta})} = 1 - [1/I(\hat{\theta})].$$

As can be seen from Table 2, the adaptive test reached reliabilities of .80, .85 and .90 after an average of only 5, 6 and 9 items respectively. In comparison to the conventional tests, the adaptive test required from 50% to 69% fewer items to reach these reliabilities. Although the adaptive test also reached reliabilities of .95 with greater efficiency than the conventional tests, the improvement in test efficiency was less than at other reliability levels (34% to 58%).

Estimated reliabilities for the adaptive and conventional tests at various fixed test lengths are given in Table 3 and are portrayed graphically in Figure 3. Note that the adaptive test is more reliable than the conventional tests at all test lengths. As expected, the differences between adaptive and conventional test reliabilities are greater at shorter test lengths.

Validity: Table 4 contains the validity coefficients between the criterion measures described in Table 1 and scores generated from the adaptive and conventional tests at various lengths. These coefficients were based on theta scores for the adaptive test and number correct scores for the conventional tests. It is apparent from this table that the differences in the validity coefficients between adaptive and conventional tests are weaker and less consistent than the differences found for test reliabilities and efficiencies. For 75% (128 out of 168) of the possible comparisons between adaptive and conventional validity coefficients presented in Table 4, the adaptive test validities matched or exceeded the conventional test validities. It is

important to note, that for the 48 possible comparisons involving the Drake and Seashore tests, the most extensively validated criterion measures, only one conventional validity coefficient (the 10-item flat test with the Seashore) was larger than the corresponding adaptive test coefficient (.73 versus .71). This difference was not statistically significant at the .05 level. Note that the validity coefficient of .70 between the Drake test and the 15-item adaptive test was not reached by either the peaked or flat test at 40 items and required 25 items to be matched by the maximum discrimination test. The validity coefficient of .75 between the Seashore test and the 20-item adaptive test was not reached by any of the other conventional tests after 40 items. Most of the differences in validities favoring the conventional tests involved the 5- or 2- point scale items from the investigator-designed questionnaire (See Table 1). The largest of these differences (.04) was between the maximum discrimination and the adaptive test at 10 items with "Musical Experience in College" as the criterion variable. This difference was only marginally significant (p=.04), and could be explained easily by sampling error given the large number of comparisons that can be made among the present validity coefficients.

As one would expect, the most sizeable increases in the validity of a given test occurs as test length increases from 5 to 10 items. The validities of both the adaptive and conventional tests are very stable after 20 items. For example, the greatest difference in the validity coefficients for a given test at 20 and 40 items is only .04. For the adaptive test this difference is only .02. In fact, the

adaptive test validities at 10 and 40 items differ by no more than .02 for all criterion measures except the Seashore test. In other words, quadrupling the length of the 10 item adaptive test does not significantly increase test validity. These results, along with the previous findings on reliability, indicate that the adaptive test provides reasonably reliable and valid scores after only 10 items.

Since the adaptive and conventional test were newly developed, the present investigators were interested in comparing the validity of these tests to those of a well-established tonal memory test like the Seashore. Consequently, in contrast to its use as a criterion variable in the previously described analyses, the Seashore test is used here as a predictor variable. Validity coefficients of the Seashore test and the adaptive and other conventional tests at 30-items are given in Table 5. Examination of Table 5 reveals that for every criterion measure, the validity of the Seashore test was lower than the validities of the other tests.

The present results are summarized as follows:

1) The 30-item adaptive test provided comparable or, in the vast majority of cases, superior measurement precision to the conventional tests at all ability levels.

2) The adaptive test required 34% to 69% fewer items to reach reliabilities comparable to the conventional tests.

3) Although differences tended to be small, in most cases, adaptive test validities exceeded those of the conventional tests.

4) The differences in reliability and validity favoring adaptive over conventional tests were strongest at short test lengths.

5) Reasonably reliable and valid adaptive test scores were achieved after only approximately 10 items.

6) Validity estimates for the investigator-designed adaptive and conventional tests exceeded those of the Seashore test on all criterion measures.

Discussion

The present results indicate that adaptive testing can provide significant improvements in test reliability, efficiency and validity in a domain where such improvements are sorely needed (see Whellams., 1971 and Vispoel, 1987 for in-depth discussions of the shortcomings of selected listening tests). Problems of poor reliability and efficiency are present even in the best existing conventional listening tests like Gordon's Musical Aptitude Profile (1965). As Whellams (1971, pp. 416) notes:

> "The norms associated with Gordon's Musical Aptitude Profile indicate that ninety-nine percent of testees in the standardization sample scored at least one-quarter to nearly one-half (depending on age) of the available points. In other words, all testees spend at least a quarter of the total test administration time, i.e., about two and half hours responding to items which play no part in discriminating between them. This means that on the average, every MAP tester has forty minutes of his time wasted."

Because of the shortcomings in commercially available music tests, many music educators have become disenchanted with them and have sought alternative methods of evaluation (Davies, 1978). Unfortunately, until the advent of computerized adaptive tests, few alternatives have emerged. Adaptive tests should be of particular interest to music educators because the most serious shortcomings of commercially

available tests (poor reliability, efficiency and validity) are the very ones that this innovative testing procedure is best at overcoming.

The improvements in test reliability, efficiency and validity obtained in the present study are similar to those found in previous research comparing adaptive and conventional tests in other content areas (e.g., Urry, 1977; Weiss, 1982 and McBride and Martin, 1983). The adaptive test of tonal memory provided superior measurement precision to the conventional tests across all ability levels except at those levels where a conventional test was peaked.

Perhaps the most dramatic differences between the adaptive and conventional tests was in their efficiencies. The adaptive test typically required one-half to two-thirds fewer items to reach levels of reliability comparable to the conventional tests. Such marked reductions in test length are welcome and could significantly reduce the fatigue and boredom effects often reported in the music test literature (e.g., McLeish, 1968). An important benefit of adaptive testing procedures is that these reductions in test length may be accomplished with no significant decrease in test reliability and validity.

In comparison to reliability and efficiency, the present findings for validity were less striking. The differences in validities between the present adaptive and conventional tests favored the adaptive test in most, but not all, cases. These inconsistent results are not without precedent. For example, in an article by McBride and Martin (1983), two validity studies are reported. In the first study,

conventional tests of 15 to 30 items had slightly higher validities than equivalent-length adaptive tests. In the second study, the adaptive test had higher validities than the conventional tests at all reported test lengths. It should be emphasized, however, that differences in validities favoring conventional over adaptive tests are rarely statistically significant. In most cases, adaptive tests meet or exceed the validities of the conventional tests. Even if conventional and adaptive tests have comparable validities, the superior reliability and efficiency of the adaptive tests frequently make them the more attractive alternative.

Additional evidence supporting the validity of the present adaptive and conventional tests was obtained by comparing the validity coefficients of these tests to those of the Seashore test, a well-established and widely used measure of musical ability. In all cases, the validities of the present tests exceeded those of the Seashore. Other weaknesses in the Seashore test revealed in the present analyses were a narrow bandwidth and a test ceiling at moderate ability levels. These weaknesses may be partly responsible for the low validities of Seashore test scores found in this study. It should be noted, however, that the reported validity and reliability results are similar to those found in the Seashore test manual (Seashore et al, 1960). These results imply that tests superior to the Seashore may be constructed using items from the present adaptive test pool.

Although the present results provide evidence supporting the feasibility of adaptive listening tests, several limitations must be

acknowledged. First, as with all simulation-type studies, the results may not parallel those found using actual data. However in a study by the present investigators (Vispoel and Twing, 1989), nearly identical results were obtained using actual and simulated data.

Second, the present results are based on the assumption that order effects do not influence responses. Recall that the examinees did not respond to the actual adaptive and conventional tests as described in this study. Instead, they responded to items from the adaptive test pool as a fixed-order conventional test. Their responses to these items then were rearranged by the computer to derive the desired adaptive and conventional test scores. It is important to remember, however, that if the Item Response Theory assumption of local independence holds, order effects should not significantly affect these responses. On the basis of results from Vispoel (1987), the present items appear to satisfy the assumptions of Item Response Theory.

Finally, the present results assume that presentation mode, paper and pencil versus computer administration, does not affect responses. According to Green et al., (1984), item parameters may be determined at first in paper and pencil mode, but must be verified when used in computer presentation. For example, there is no guarantee that item difficulties will not change. If such differences exist, ability scores from computer adaptive test administration can be equated to the paper and pencil scores. Presentation mode also may influence the validity of test scores. However, the limited research that has been done (McBride, 1980; Sympson, et al, 1982) indicates that presentation

mode has little effect on validity. If these results generalize to other situations, then "post hoc" validity analysis of the type described in this study would seem to be worthwhile due, in part, to the resources needed to conduct full-scale adaptive test validity studies.

Validating an adaptive test is a complicated and timeconsuming process. It often is difficult to derive the item parameter estimates for potential adaptive test items and to obtain enough items with which to measure all ability levels with high precision. According to Lord (1980), sample sizes of 1000 or more examinees are necessary for obtaining stable estimates for the item parameters in the three parameter logistic model. In addition, it is difficult to find examinees willing to respond to the large number of items required to reliably and efficiently assess all levels of examinee ability. In calibrating items, it is not unusual to eliminate as many as 50% of them due to their poor quality. Consequently, several administrations of pilot items may be necessary to obtain enough items to fill the adaptive test pool. Once an adequate item pool is obtained and an operational the adaptive test is constructed, it is still necessary to validate the test on a new sample of examinees. Since the adaptive test is administered at a computer terminal it is usually difficult to obtain validation data on a large number of examinees because the test only can be administered by to one examinee at a time. This problem could be alleviated to a certain extent if many terminals were available, but even under the best of circumstances large amounts of

validity data are difficult to obtain using computerized adaptive tests. This may explain why few, if any, existing adaptive tests are validated adequately.

The problems in thoroughly validating adaptive tests can be reduced to a significant degree if, in fact, "post hoc" validation studies are indicative of results from subsequent full-scale validation studies. In a "post hoc" validity study, criterion measure scores and item responses to potential adaptive test items would be derived for the original calibration sample of examinees. Through computer simulations, the efficiency, reliability and validity of investigator-designed adaptive and conventional tests can be assessed and compared to one another using procedures similar to those used in the present study. Such procedures would enable one to assess the overall quality of a computerized adaptive test and determine if it is any more useful than a conventional test in meeting the needs of the test user. For example, if concurrent validity was the main criterion for test selection in this study, a 20-item conventional test would yield validity estimates similar to the adaptive test without the added time, effort and expense involved in building an adaptive test. To provide a mechanism for evaluating adaptive tests, the present authors have developed a computer program to compare the efficiency, reliability and validity of computerized adaptive tests with any potential conventional test generated from an item pool constructed by the user (Twing & Vispoel, 1989).

Conclusion

As the popularity of computerized adaptive tests grows, the need to assess the quality of such tests will become increasingly important. If the present results for adaptive listening tests are any indication, adaptive testing procedures will be applicable to a wide range of disciplines. A simulation program like the one employed in this study is an important first step in developing efficient and cost-effective procedures for assessing the reliability, efficiency, and validity of these tests.

REFERENCES

Bentley, A. (1966). Measures of musical abilities, manual. London: Harrap.

Crichton, L. C. (1981). Effect of error in item parameter estimates on adaptive testing. Unpublished doctoral dissertation, University of Minnesota.

Davies, J. B. (1978). The psychology of music. London: Hutchinson.

Gordon, E. (1965). Musical aptitude profile. Boston: Houghton Mifflin.

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L. & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. Journal of Educational Measurement, 21, 347-375.

Haebara, T. (1980). EQUATOR: A program for equating logistic ability scales. Unpublished computer program, University of Iowa.

Hambleton, R. K. & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer Nijhoff Publishing.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Applications to psychological measurement. Homewood, IL: Dow JonesIrwin.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum & Associates.

McBride, J. R. (1980). Adaptive verbal ability testing in a military setting. In D. J. Weiss (Ed.), Proceedings of the 1979 Computerized Adaptive Testing Conference. Minneapolis: University of Minnesota, Department of Psychology.

McBride, J. R. & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic Press.

McLeish, J. (1968). Musical cognition, Research Paper No. 2. London: Novello.

Seashore, C. E., Lewis, D. & Saetveit, J. C. (1960). Manual of instructions and interpretations for the Seashore measures of musical talents (2nd rev.). New York: The Psychological Corporation.

State of Illinois. (1985) Public Act 84-126. State of Illinois, Springfield, ILL.

Sympson, J. B., Weiss, D. J., & Ree, M. (1982). Predictive validity of conventional and adaptive tests in an Air Force training environment (AF HRL-TR-81-40). Brooks Air Force Base, TX: USAF Human Resources Laboratory.

Twing, J. S., & Vispoel, W. P. (1989). SIMUCAT: A program to simulate and score Item Response Theory based adaptive and conventional tests. Paper presented at the 31st international ADCIS Conference, Washington, D. C.

Urry, V. W. (1977). Tailored testing: A successful application of latent trait theory. Journal of Educational Measurement, 14 (2), 181-196.

Vispoel, W. P. (1987). An adaptive test of musical memory: An application of item response theory to the assessment of musical ability. Unpublished doctoral dissertation, The University of Illinois: Urbana -Champaign.

Vispoel, W. P., & Twing, J. S. (1989). Creating adaptive tests of musical ability with limited size item pools. Paper submitted to the 31st International ADCIS Conference, Washington, D. C.

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. Applied Psychological Measurement, 6, 473-492.

Whellams, F. (1971). The aural musical abilities of junior school children: A factorial investigation. Unpublished doctoral dissertation, University of London.

Wing, P. D. (1968). Tests of musical ability and appreciation (2nd ed.). British Journal of Psychology Monograph Supplement.

Wingersky, M. S., Barton, M. A. & Lord, F. M. (1982). LOGIST user's guide. LOGIST V, version 1.0. Princeton, NJ: Educational Testing Service.

Table 1
A Description of Criterion Measures.

MEASURE	DESCRIPTION
Drake Musical Memory Test	A 54-item standardized test of musical memory. (In this study, Mean = 32, SD = 7.3, KR20 = .82, N = 250).
Seashore Tonal Memory Test	A 30-item standardized test of tonal memory. (In this study Mean = 25.2, SD = 4.6, KR20 = .86, N = 253).
Instrument Playing Experience	Examinees indicated the number of years they have played a musical instrument (Mean = 7.2, SD = 7.4, N = 336).
Perceived Musical Ability	Examinees rated their overall musical ability on a five-point scale where 1 = low ability, and 5 = high ability (Mean = 3, SD = 1.25, N = 336).
Perceived Ability to Tune an Instrument.	Examinees rated their overall ability to tune an instrument on a five-point scale where 1 = low ability, and 5 = high ability (Mean = 2.4, SD = 1.48, N = 336).
Professional Musical Performing Experience.	Examinees indicated whether they had ever played music professionally (been paid to perform music). They were coded 1 if they had played professionally and 0 otherwise (Mean = 0.13, SD = 0.34, N = 336).
College Music Experience	Examinees indicated whether they were music majors or music minors. They were coded 1 if they were a music major and 0 otherwise (Mean = 0.13, SD = 0.34, N = 336).

Table 2
Comparisons of the Average Test Lengths Needed to Obtain Selected Reliability Estimates for the Adaptive and Conventional Tests.

Characteristic	Test(s)	Reliability Estimates			
		.80	.85	.90	.95
Average Test Length	Adaptive	5	6	9	19
	Peaked	16	18	21	29
	Flat	14	16	23	46
	Max. Disc.	13	13	18	33
Differences in Test Length	Peaked - Adaptive	8	7	9	14
	Flat - Adaptive	11	12	12	11
	Max. Disc. - Adaptive	9	10	14	27
Percent Reduction in Test Length using the Adaptive Test.	Peaked	69	67	57	34
	Flat	64	63	61	58
	Max. Disc.	62	54	50	42

Table 3
Estimated Reliabilities at Various Test Lengths.

Test Length	Adaptive	Peaked	Flat	Maximum Discimination
5	.80	.00	.29	.00
10	.91	.28	.74	.75
15	.94	.75	.83	.88
20	.96	.88	.88	.91
25	.96	.93	.91	.93
30	.97	.96	.93	.95
35	.97	.96	.94	.96
40	.98	.97	.94	.96

Table 4
Correlations Between Criterion Measure Scores
and Four Adaptive and Conventional Tests.

Criterion	Test	Test Length 5**	10	15	20	25	30	35	40
Drake Test (N=249)	Adaptive	.57	.68	.70	.70	.70	.70	.70	.70
	Peaked	.51	.62*	.64*	.65*	.66	.67	.67	.67
	Flat	.47*	.64	.65*	.66	.66	.68	.68	.69
	Max Disc.	.56	.65	.68	.69	.70	.70	.70	.70
Seashore Test (N=249)	Adaptive	.65	.71	.73	.75	.76	.76	.77	.77
	Peaked	.59	.66	.67*	.70*	.71*	.71*	.72*	.72*
	Flat	.61	.73	.73	.72	.72*	.73	.73*	.73*
	Max Disc.	.51*	.61*	.69	.68*	.72*	.70*	.70*	.70*
Instrument Playing Experience (N=334)	Adaptive	.48	.50	.49	.50	.50	.51	.51	.50
	Peaked	.42	.48	.48	.49	.50	.51	.51	.52
	Flat	.38*	.43*	.47	.49	.50	.51	.51	.52
	Max Disc.	.40*	.47	.49	.50	.49	.51	.52	.51
Perceived Musical Ability (N=334)	Adaptive	.49	.51	.52	.51	.52	.53	.52	.52
	Peaked	.51	.54	.54	.55	.55	.56	.56*	.56*
	Flat	.39*	.45*	.50	.51	.53	.53	.53	.54
	Max Disc.	.45	.51	.54	.53	.54	.54	.54	.55
Perceived Ability to Tune an Ins. (N=334)	Adaptive	.49	.52	.53	.52	.53	.53	.52	.52
	Peaked	.45	.50	.52	.52	.53	.53	.53	.54
	Flat	.44	.48	.50	.51	.54	.53	.54	.55
	Max Disc.	.44	.50	.52	.52	.53	.53	.53	.54
Professional Performing Experience (N=334)	Adaptive	.42	.44	.45	.47	.47	.48	.47	.47
	Peaked	.39	.45	.46	.46	.46	.47	.47	.47
	Flat	.37	.42	.43	.45	.46	.46	.47	.48
	Max Disc.	.41	.47	.45	.47	.46	.48	.49	.49
College Music Experience (N=334)	Adaptive	.47	.48	.48	.49	.50	.50	.50	.50
	Peaked	.42	.46	.47	.48	.48	.49	.49	.49
	Flat	.39	.45	.46	.48	.49	.50	.51	.52
	Max Disc.	.49	.52	.50	.52	.50	.53	.54*	.54*

* Indicates a significant difference between the adaptive and conventional test correlations ($p<.05$).

** For five item tests, N=233 for Seashore and Drake.

Table 5
Validity Coefficients for the 30 Item Adaptive and Conventional Tests (n = 248).

Criterion Measure	Seashore	Adaptive	Peaked	Flat	Maximum Discrimination
Drake	.65	.70	.66	.67	.70*
Instrument Playing Experience	.27	.36*	.37*	.38*	.35*
Perceived Musical Ability	.38	.44	.47*	.46*	.46*
Perceived Ability to Tune an Instr.	.34	.43*	.42*	.43*	.44*
Professional Playing Exp.	.21	.36*	.34*	.35*	.35*
College Music Experience	.17	.32*	.29*	.33*	.34*

* Indicates a significant difference ($p<.05$) between the Seashore and given test validity coefficients.

Figure 1. Plots of Information.

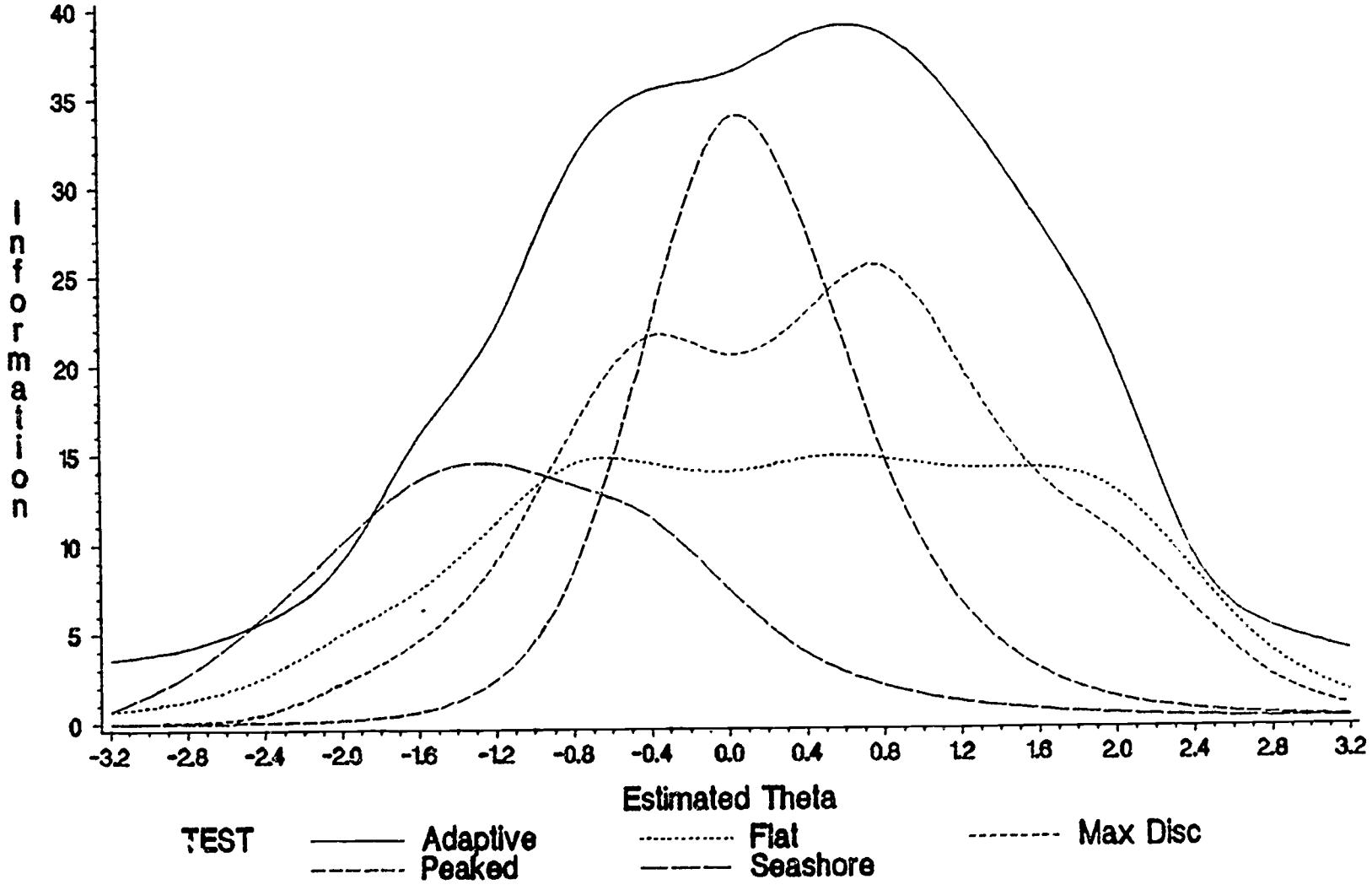

Figure 2. Plots of Relative Efficiency: Adaptive Test as the Standard.

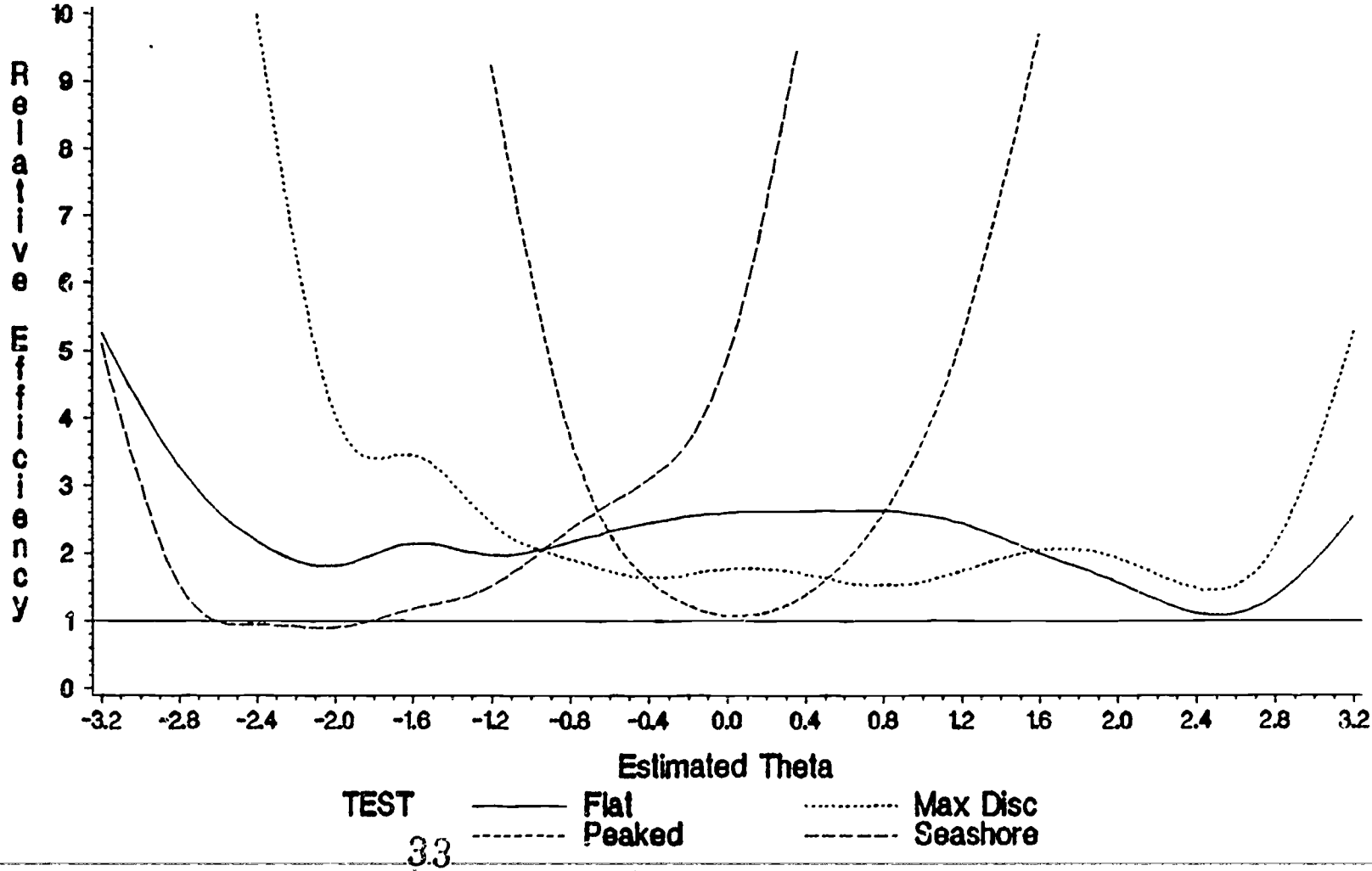

Figure 3. Plots of the Average Number of Items as a Function of Reliability.