In this study, I investigated the effects of four score study styles—no score study, study with score alone, study with score and correct aural example, and score study at the keyboard—on the pitch and rhythm error-detection abilities of beginning conducting students. Four computer-assisted tests were developed, each with 31 examples using sampled sounds and ranging from one to eight parts. A counterbalanced design was used to control for effects of presentation order as to score study style and individual test. Score study with a correct aural example was found to be significantly more effective than study with the score alone. No significant difference was found between score study at the keyboard and any other score study style. There were significant differences in test scores attributable to the number of parts in examples. Generally, error detection became more difficult as the number of parts increased.

Don R. Crowe, South Dakota State University

Effects of Score Study Style on Beginning Conductors’ Error-Detection Abilities

The ability to detect errors in performance is a basic skill for all conductors. If the director of an ensemble cannot readily discern when and where errors are being made, rehearsal time becomes less effective and the ensemble’s level of performance suffers. In a survey of music teachers, music consultants, and university music faculty, Taebel (1980) found that error-detection skills were considered by music teachers to be among the most important competencies used in their teaching.

Investigators of error detection have considered both the relationship of error-detection ability to other factors and the effects of specific methods of training on error-detection ability. These studies have involved a wide range of subjects, error types, excerpt types, presentation media, and environments.

The preponderant evidence from investigations of the first type indicates that no factor other than experience is related to error-detection ability (Blocher, 1986; Boyer, 1974; Brand & Burnsed, 1981; Byo, 1993; Collings, 1973; Gonzo, 1970; Shaw, 1971). Detection of errors seems to be more difficult in multitimbral music as opposed to music with only one timbre, in multipart music as opposed to solo music regardless of timbral considerations, and in excerpts with more than

Don R. Crowe is an assistant professor of music education in the Department of Music at South Dakota State University, Brookings, SD 57007. Copyright © 1996 by Music Educators National Conference.
one error (Byo, 1993; Mount, 1982). A number of researchers (Boyer, 1974; Brand & Burnsed, 1981; Doane, 1989; Gonzo, 1970; Jones, 1990; Larson, 1977; Shaw, 1971; Sidnell, 1971; Simard, 1982) have concluded that instruction in ear training as it is traditionally presented in colleges and universities is inadequate to prepare future educators for rehearsal and performance situations in which they must detect and correct errors in music while reading a score with multiple staves and dealing with many different timbres.

Studies concerning the effects of training on error-detection skills (Behmer, 1984; Boyer, 1974; Collings, 1973; Costanza, 1971; Deal, 1983; DeCarbo, 1981; Doane, 1989; Dolbeer, 1969; Gruner, 1993; Grunow, 1980; Hopkins, 1991; Jones, 1990; Malone, 1985; Michels, 1973; Ramsey, 1979; Rizzolo, 1969; Shaw, 1971; Sidnell, 1968; Stuart, 1978; Stwolinski, Faulconer, & Schwarzkopf, 1988; Tromblee, 1973) have used live, taped, and computer-generated performances. Instruction in error detection has been presented as a part of a regular class, in a self-instructional format, and combined with conducting experiences and with instruction in other generic teaching skills. Subjects have ranged from high school to doctoral students and experienced teachers and have been drawn from both heterogeneous groups such as conducting classes and groups of specifically choral or instrumental students.

Several researchers have focused on computer-assisted training in error detection. Deal (1983) compared the effects of a programmed instruction approach developed by Ramsey (1979) with those of a computer-assisted program based on it. Deal's program featured four-voice excerpts in three timbres. Both programs resulted in significant gains in error-detection ability.

Jones (1990) tested a computer-assisted error-detection instruction program for undergraduate and graduate instrumental music education students. He concluded that the program, which featured full-band excerpts presented in a printed score booklet, was effective in developing the subjects' abilities to detect errors of pitch, rhythm, articulation or style, and ensemble.

Gruner (1993) investigated the effects of a computer-assisted training program, which featured synthesized wind instrument timbres in two- to five-part excerpts taken from band literature of varying difficulty, on the pitch and rhythm error-detection abilities of undergraduate instrumental music education majors. The scores were presented in printed booklets. He found that the treatment group improved significantly in overall error-detection ability, that both the treatment and control groups improved significantly in rhythm error detection, and that neither group improved significantly in pitch error detection.

Several studies have investigated the effects of score study styles on error-detection ability. Collings (1973) found no significant difference in the effects of three error-detection training techniques—no prestudy, prestudy followed by error examples, and prestudy followed by correct examples—on the pitch error-detection abilities of second semester undergraduate conducting students. Grunow (1980) found no significant difference between four modes of score study—score
alone, score plus a recorded example, recorded example alone, or no study—on the development of "visual-aural discrimination skills." Hopkins (1991) found that score study with a correct recorded example was significantly more effective than score study at the piano in preparing undergraduate music students to detect pitch and rhythm errors in complete choral works. Stwolinski, Faulconer, and Schwarzkopf (1988) found that listening to a correct rendition of simple piano compositions better prepared piano class students to detect errors in those works than did sight-reading at the piano.

The wide variety of subjects, conditions, training procedures, and error types featured in error-detection studies and the recommendations made by the investigators indicate the need for further research. These recommendations include the use of multiple timbres (Byo, 1993; Collings, 1973; Sidnell, 1968), ensembles of various sizes (Collings, 1973; Deal, 1983; Gruner, 1993; Grunow, 1980; Jones, 1990), and scores in written pitch (Collings, 1973) with an increasing number of parts and timbres (Ramsey, 1979; Sidnell, 1968; Stuart, 1978). A focus on errors of pitch and rhythm only (Collings, 1973; Stuart, 1978) in clear performances with preplanned errors (Dolbeer, 1969; Gruner, 1993) is suggested, along with score study at the piano (Dolbeer, 1969). Finally, an investigation of computer-assisted instruction (Behmer, 1984; Boyer, 1974) using computer-generated timbres (Deal, 1983; Gruner, 1993) with scores presented on the computer screen (Gruner, 1993) is recommended.

A search of the literature yields no evidence that a comparison between the effects of score study at the keyboard and other methods of score study on the ability to detect errors in instrumental performance has been attempted. The technology now exists that permits investigators to incorporate computer-assisted presentation of materials and scoring of subject responses, a variety of sampled instrumental timbres with scores in written pitch, a varied and increasing number of parts in error examples, and the use of preplanned errors of pitch and rhythm only in "clean" performances.

It was the purpose of this study to investigate the effects of four different modes of score study on the abilities of undergraduate conducting students to detect errors of pitch and rhythm in four- to six-measure excerpts of from one to eight parts drawn from band literature. The four modes of score study were: no prestudy (X1); study with the score alone (X2); study with the score and a correct aural example (X3); and score study at the keyboard (X4).

**METHOD**

Subjects for the study were 30 undergraduate students, members of beginning conducting classes at three midwestern universities. The study took place during the fall semester of 1993.

I rescored error examples for the study from full-score examples developed for an earlier study of error detection by Ramsey (1979); the examples varied from four to six measures in length and from one to
eight parts. The pitch and rhythm errors were not changed from the original, although the original scores were condensed through the elimination of parts. The examples were assigned to four sets, using a modified random assignment to ensure equivalence between sets, and rescored into a specific number of parts according to the difficulty of the originals, as reported by Ramsey. The most difficult were given eight parts. There were 12 examples with eight parts, three in each set (one brass and two woodwind), and 16 examples each of from one to seven parts, four in each set (two brass and two woodwind).

A set of six Hypercard (Atkinson, 1987–90) stacks was developed for the presentation of these examples and the management of the study: an introductory stack, four example set stacks, and a statistics stack. Using a counterbalanced design, the four possible score study styles were entered into a Latin square. I treated the “no prestudy” condition as part of the counterbalanced design rather than as a control or baseline—which would have required a separate, initial testing—so as to take full advantage of the research design in controlling for the possible contaminating effects of repeated testing, maturation, practice, transfer, and history on subjects’ achievement (Campbell & Stanley, 1963).

As shown in Table 1, the introductory stack assigned score study style orders on a rotating basis to each subject in a group upon his or her first session. Each subsequent subject was assigned a different order until all four had been assigned, and then the series was repeated. This stack also initiated the correct score study style for each session according to the order assigned to the subject, and opened the correct example set stack for each session. The order of example sets was the same for every subject in any one group, but different for each group through the application of another Latin square. Over the course of four sessions, a subject would receive every score study style and every

---

Table 1

<table>
<thead>
<tr>
<th>Student</th>
<th>Session</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>1, 5, ...</td>
<td>X1</td>
</tr>
<tr>
<td>2, 6, 10...</td>
<td>X2</td>
</tr>
<tr>
<td>3, 7, 11...</td>
<td>X3</td>
</tr>
<tr>
<td>4, 8, 12...</td>
<td>X4</td>
</tr>
</tbody>
</table>

*Note. X1 = no score study
X2 = study with score alone
X3 = study with score plus correct example
X4 = study with score at the keyboard*
example set. Twenty-one subjects completed all four sessions; the data from all completed sessions were included in the statistical analyses that used score study style and example set as independent variables, but data for subjects who did not complete all four sessions were not included in the analyses, which used score study style order or example set order as independent variables.

Each example set stack was capable of presenting the materials under every score study condition. Each opened with instructions for the particular score study style assigned for the session and a practice example. The 31 test examples in the set were then presented, and the subject was required to use the computer's mouse to point to and click on the part, the measure, and finally the specific note that was incorrect. The program recorded the number of attempted answers, as determined by the number of mouse clicks, taken for each example, as well as the total time spent on each example and the amount of time spent studying each. Upon successfully completing an example, or having reached the limit of allowed answer attempts (10 per example), the subject was shown the error as it was played before the next example was presented. When the example set was complete, the data recorded during the session were copied to a statistics stack and the example set stack was readied for the next subject.

Data were analyzed to determine whether significant differences occurred in test scores that could be attributed to (a) score study style, (b) order of presentation of score study styles, (c) example set/individual test, (d) intact group/order of presentation of example sets, or (e) the number of parts in the examples.

RESULTS

The content validity of the example sets/tests was inferred from the validation procedures used by Ramsey (1979), as all excerpts were drawn from his program and test. In his study, instrumental music teachers identified typical or often-corrected errors of pitch and rhythm. These were inserted into excerpts from commercially available band literature of medium difficulty, with care taken to avoid altering the music in any way other than the desired error. A panel of experts evaluated recordings of these error excerpts to determine whether the errors were typical and to validate the recordings themselves. Finally, a pilot test was undertaken with a group of instrumental music majors in order to develop a continuum of difficulty using item analysis. The validated errors were not changed for the purposes of this study; the number of parts surrounding each error part were controlled as described above.

As a check of the reliability of the example sets used in this study, one set was chosen at random and tested for internal consistency using a split-half approach. A Pearson product-moment correlation was calculated for Example Set D and corrected using the Spearman-Brown prophecy formula. A reliability of .81 was established for this set.

Error-detection ability was measured by the total number of attempts
made in answering the error identification questions per session and the number of such attempts per excerpt. The total time taken per session and score study time taken per session were considered to be functions of the score study style assigned. Means were calculated for each of these measures and tabulated by subject, group, score study style, example set, session, and number of parts in excerpts. These data were then analyzed using one-way analyses of variance. Where indicated, multiple paired comparisons were made using the Newman-Keuls procedure.

Significant differences were found on test scores attributable to score study style \( F(3, 93) = 2.929, p < .05 \). Score study with a correct aural example was found to be significantly more effective, in terms of test scores, than study with the score alone \( (p < .05) \). No significant difference was found between score study at the piano and any other score study style.

There were significant differences in test scores attributable to the number of parts (staves) in the examples \( F(7, 768) = 67.826, p < .01 \). Significant differences existed between every pair of part numbers except examples with two and those with four parts, examples with three and those with four parts, and examples with five and those with six parts. Mean attempts were lowest for examples with one part \( (M = 6.044, SD = 0.830) \) and highest for those with eight parts \( (M = 9.478, SD = 1.415) \), and generally increased from one level to the next. The only exception was a drop from three-part \( (M = 7.773, SD = 1.276) \) to four-part \( (M = 7.649, SD = 1.017) \) examples.

No significant differences in test scores were found attributable to the order of presentation of score study styles \( F(3,12) = 1.951, p > .05 \), the individual example set/test \( F(3, 93) = 1.800, p > .05 \), or the order of presentation of the example sets/intact group \( F(2,94) = 0.050, p > .05 \).

In order to validate the procedures used in the study, analysis of mean score study time and mean total time taken per session was undertaken with regard to the effects of score study style, example set, and session number. It was assumed that differences found in score study time per session that were attributable to score study style and that followed the trend \( \text{Time X4} > \text{Time X3} > \text{Time X2} > \text{Time X1} \) would indicate that subjects were indeed using the styles prescribed in each session. The raw data did reflect this trend.

There were significant differences between mean score study time per session, measured in seconds, attributable to score study style \( F(3, 93) = 4.837, p < .01 \). These occurred between the “no score study” approach and both the “score with correct aural example” and “score study at the keyboard” approaches. Mean score study time per session in seconds was lowest under the “no score study” condition \( (M = 4.833, SD = 2.078) \) and highest under the “score study at the keyboard” condition \( (M = 356.522, SD = 630.565) \).

There was a significant difference in total time per session, measured in seconds, attributable to session number \( F(3, 93) = 7.153, p < .01 \). Significant differences were found between the first session and each of
the others. Mean total time per session dropped with each subsequent session, being highest for Session 1 (M = 1445.100, SD = 755.124) and lowest for Session 4 (M = 872.905, SD = 265.926). No significant differences were found in total time per session attributable to score study style \([F(3, 93) = 1.239, p > .05]\), in either score study time \([F(3, 93) = 0.969, p > .05]\) or total time \([F(3, 93) = 1.166, p > .05]\) attributable to example set, or in score study time attributable to session number \([F(3, 93) = 1.826, p > .05]\).

In summary, significant differences in error-detection test scores occurred that were attributable to (a) the score study style implemented in test sessions and (b) the number of parts in the examples. Score study with a correct aural example was found to be significantly more effective than study with the score alone. As to the number of parts in examples, the results of post hoc \(t\)-tests show significant differences between every pair of part groups except those with two and those with four parts, examples with three and those with four parts, and examples with five and those with six parts. Generally, error detection became increasingly more difficult as the number of parts in the excerpts rose, regardless of score study style, example set, or session. Significant differences were also found in mean score study time per session attributable to score study style, and in mean total time per session attributable to session number.

**DISCUSSION**

Study of the score with a correct aural example was found to be significantly more effective than study with the score alone in preparing subjects to detect errors of pitch and rhythm. These results differ in some respects from the findings of other investigations of the effects of score study style on error-detection ability. The differences may be attributable in part to the differences in error types, excerpt types, procedures, and subjects involved in the studies.

Collings (1973) found no significant difference in the error-detection abilities of second-semester undergraduate conducting students following training using three methods of score study: no study, score study followed by error examples, and score study followed by correct examples. The subjects in his study received six 30-minute instructional sessions, and only pitch errors were used, with from one to five being randomly inserted into brass quintet and sextet excerpts of six measures' length. Excerpts used on the pretest and posttest were randomly arranged with regard to the number of errors in each.

Grunow (1980) investigated the effects on error-detection ability of training featuring four modes of score study—no study, study with the score alone, study with score and a correct recorded example, and study with the recorded example alone—and found no significant differences in the developed error-detection abilities. The errors investigated were placed into one of two categories. “General Musical Criteria” included tempo, balance, style of articulation, tone quality, and intonation, while “Specific Technical Criteria” included note accu-
racy, pitch accuracy, phrasing, dynamic contrast, and ensemble. The low reliability of the "General Musical Criteria" portion of the test ($r = .40$) and of the overall test ($r = .66$) call these results into question.

Stwolinski, Faulconer, & Schwarzkopf (1988) investigated the differential effects of keyboard sight-reading and listening to recordings on the abilities of undergraduate music students to detect harmonic errors in piano excerpts. The subjects in this study had "demonstrated minimum keyboard proficiency in the third semester of college-level class piano" (p. 84). Four excerpts of from 6 to 12 measures' length and of equivalent difficulty were chosen, and four to five harmonic alterations were inserted into each. All four excerpts were completed within 40 minutes. Whereas in the present study subjects either (a) heard the correct rendition only once or (b) played the excerpt in any manner desired as many times as desired before hearing the single presentation of the error excerpt, Stwolinski's subjects either (a) heard three correct repetitions of each excerpt or (b) sight-read the excerpt at an electronic piano three times in succession before hearing the single presentation of the error excerpt. Subjects then marked the printed score on the beats containing perceived errors. Both missed errors and incorrectly identified beats were considered in scoring. The listening condition was found to be significantly more effective in preparing the subjects for the error-detection task than the sight-reading condition. The discrepancy between these findings and those of the present study may be attributable to the differences in procedures noted above, or to differences in the subject populations (third-semester piano class students versus undergraduate conducting students).

Hopkins (1991) compared the effects of score study with the score alone, with the score at the piano, with the score and a correct recorded example, and sight-singing from the score on the error-detection abilities of undergraduate choral music education students. The subjects completed one score study session under each condition, with a 50-minute time period allotted for studying the full score of a four-part motet. Each error excerpt contained five pitch and five rhythm errors and was presented three times. The subjects were to both detect and notate the errors. Score study with a correct recorded example was found to be significantly better than score study at the piano, but no other significant differences were identified. The differences between these findings and those of the present study may be attributable to the differences in materials (whole choral compositions versus short wind-band excerpts), procedures (50-minute score study periods, multiple errors, and repeated hearings of error excerpts versus open-ended but short score study time, single errors, and single presentations of error excerpts), and subjects (choral music education students versus heterogeneous groups of undergraduate conducting students) incorporated in the studies.

Two of the above studies investigated error-detection training and used excerpts with either multiple errors of pitch alone or multiple error types artificially combined into two categories. Another compared the effects of class piano students’ sight-reading of piano
excerpts with repeated hearings of correct versions of those excerpts. The last, while using a one-shot approach to score study as did the present investigation, allowed 50 minutes for the study of a single whole work of choral music and used multiple errors and multiple presentations of error excerpts. Further investigation into this area is needed to resolve the contradictions between the findings of these studies and those of the present one.

No significant differences were found in test scores attributable to the order of presentation of score study styles, the individual example set/test, or the intact group/order of presentation of example sets. This indicates that the tests were equivalent and that no one order of presentation of either score study styles or example sets resulted in better test scores. The lack of significant differences attributable to intact groups seems to indicate that institutional differences had no effect on test scores.

There was a significant difference between test scores on examples due to the number of parts in the examples. The results of post hoc $t$-tests show significant differences between every pair of part groups except those with two and four parts, those with three and four parts, and those with five and six parts. The test scores on this measure show that subjects found the error-detection task increasingly difficult as the number of parts increased, regardless of score study style, example set, or session.

These results corroborate the findings of the only other study that has investigated the effects of the number of parts in musical excerpts on error-detection ability. Mount (1982) investigated the various effects of listening to one voice part, paired parts, and all four parts of a Bach chorale on error-detection accuracy among graduate choral music and church music majors. Twenty-five errors were inserted into the chorale, and one phrase was presented at a time. Mount found that there were significant differences in error-detection test scores between (a) the "parts alone" and "paired parts" listening conditions, (b) the "parts alone" and "all voices" conditions, and (c) the "paired parts" and "all voices" conditions.

Any differences in score study time or total time found attributable to example set would have seemed to indicate that the four sets were not of equal difficulty; no such differences were found. The significant differences found in total time taken per session attributable to session number could have resulted from many different factors, including familiarity with the program procedures, subjects’ tiring of the program, and increasing facility in the error-detection task due to prior sessions.

In this study, score study with a correct aural example was found to be significantly more effective than study with the score alone. The example sets developed for the study were considered valid, reliable, and equivalent. There was no significant difference in test scores attributable to intact group. Finally, the number of parts in each example did have an effect on the difficulty of the error-detection task. The detection of errors became increasingly difficult as parts were added to
excerpts.

Further research into error-detection ability is indicated. The interaction between score study style and the complexity, length, and error type of error examples should be investigated, as should the comparative effectiveness of various score study styles as instructional materials. An investigation of the effectiveness of different score study styles with subjects of different experience levels, e.g., graduate students and experienced teachers, is also recommended.

A number of implications for the training of error detection arise from the results of this study. Error examples should include a variety of timbres, both as solo excerpts and in natural combinations. Increased difficulty can be provided by increasing the number of parts in error examples. Error-detection training should incorporate correct aural examples in order to develop score-reading and audiation skills. Excerpts without these examples should gradually be introduced, to encourage the development of students' aural imaging.

REFERENCES


Doane, C. (1989). The development and testing of a program for the develop-


Submitted June 14, 1994; accepted February 14, 1995.