Research has shown that undergraduate students' self-evaluation of performance correlates poorly with instructor and peer evaluation. This article reports two exploratory investigations into the development of a treatment condition for improving performance self-evaluation. The condition consisted of small groups of peers informally discussing performances and sharing feedback with one another.

The first investigation resulted in a statistically significant difference between experimental and control groups in ability to self-evaluate, although the effect size was small. With a second investigation, we pursued a modified version of the treatment emphasizing changes over time in ability to self-evaluate. In the second investigation, we also examined different effects of this modified treatment condition on students whose initial attempt to self-evaluate was either accurate or inaccurate. The second investigation did not result in significant differences between treatment and control groups; however, a significant interaction between time (self-evaluation across five small-group peer-interaction sessions) and initial ability to self-evaluate accurately was noted. A prompt improvement was found with performers whose initial ability to self-evaluate accurately was poor, although the effect tended to fade over time.

Consistent with prior research, self-evaluation did not correlate highly with instructor evaluation. Also consistent with prior research, peer evaluation was higher than instructor evaluation. Correlations between instructor evaluation and peer evaluation declined over the five sessions. Self-evaluation scores increased over time, moving away from instructor evaluation scores and toward the higher peer-evaluation scores.

Martin J. Bergee, *University of Missouri–Columbia*
Lecia Cecconi-Roberts, *Columbia, Missouri, Public Schools*

Effects of Small-Group Peer Interaction on Self-Evaluation of Music Performance

Self-evaluation has long engaged the attention of researchers. In a meta-analysis of student self-assessment in higher education, Falchikov and Boud (1989) examined dozens of studies spanning more than a half-century. Self-evaluation's potential to encourage

Martin J. Bergee is an associate professor of music education in the School of Music, 138 Fine Arts, University of Missouri–Columbia, Columbia, MO 65211; e-mail: bergeem@missouri.edu. Lecia Cecconi-Roberts is an instrumental and general music specialist in the Columbia, Missouri, Public Schools, Hickman High School, 1104 North Providence, Columbia, MO 65201.

individuals to take more responsibility for their own learning lends credence to such intensive research effort. Accurate self-assessment has led to improved job performance (Randall, Ferguson, & Patterson, 2000) and, among students, to more realistic exam preparation (Balch, 1998). Atwater and Yammarino (1996) have cautioned that negative organizational consequences could result if employees do not see themselves as others do, especially if self-ratings are inflated relative to others' ratings.

Because of its potential to bring about positive change in students, researchers have studied self-evaluation in a number of educational settings, for example, management (Walker & Warhurst, 2000), language acquisition (Ross, 1998), and medicine (Fitzgerald, Gruppen, & White, 2000). It also has been studied in music teaching contexts. Guided self-analyses have resulted in desired changes in behavior of music teachers (e.g., Arnold, 1995; Cassidy, 1993; Yarbrough, 1987) and applied music instructors (Benson, 1989). Two pedagogical approaches in particular—peer interaction/feedback and observation of videotape—have enhanced students' ability to self-evaluate music teaching skills (Colwell, 1995; Rosenthal, 1985).

Ability to self-evaluate music performance has received less attention, perhaps owing to the relatively traditional instruction most students receive in applied study. A master-apprentice relationship continues to predominate. Such techniques as small-group instruction (Corder, 1979) and peer evaluation (Byo, 1990) have rarely been used (Hepler, 1987). The majority (72%) of Daniels's (2001) respondents (undergraduate music students) indicated that instructors required them to evaluate their own performance "sometimes" to "not at all," suggesting a heavy reliance on instructor-centered evaluation.

To determine the extent to which undergraduate students were able to evaluate their own performance accurately, Bergee (1993) compared faculty, peer, and self-evaluation of end-of-semester performances of music and music education students studying brass instruments. Correlations between faculty and peer evaluation were high, ranging from .86 to .91 ($p < .01$). Consistent with earlier research on peer and self-evaluation, peer evaluation generally was higher than instructor evaluation, and self-evaluation correlated poorly with both faculty and peer evaluation. Self-evaluation showed no consistent pattern of being higher or lower than others' evaluations. Interjudge reliability among faculty and peer evaluations was high, with total score correlations ranging from .83 to .89 ($p < .01$). Faculty and peer interjudge reliability also was high on three of four subscales: interpretation/musical effect (.80–.94), tone quality (.83–.95), and technique (.74–.97). The fourth subscale, rhythm/tempo, had mixed results (.13–.81).

A second investigation (Bergee, 1997) generalized the findings of the investigation just discussed to undergraduate performers of percussion, woodwind, brass (again), stringed instruments, and voice. Consistent with results of the first investigation, correlations between

faculty and peer evaluations generally were high, ranging from .61 ($p < .10$) to .98 ($p < .01$). Also consistent with results of the first investigation, peer evaluation generally was higher than faculty evaluation, and self-evaluation correlated poorly with both faculty and peer evaluation. No significant differences in self-evaluation ability were found among performance concentrations (voice, percussion, etc.) or between lower-level (first/second year) and upper-level (third year and beyond) student performance. Ranging from .23 to .93, total score reliability of faculty evaluation panels was mixed. Total score interjudge reliability among student (peer) panels was more consistent (.83–.89). Most category (e.g., tone, intonation, etc.) score reliabilities were acceptable, although there was a wide range.

The results of these studies established that undergraduate performers were consistently unable to self-evaluate accurately. Applied music teachers, ensemble conductors, and others likely assume that undergraduate students hear their own performing the same way their instructors do. Evidence, however, suggests that students do not. Modifications to standard pedagogical approaches concerning performance assessment therefore should be considered. To develop a knowledge base of techniques of demonstrated effectiveness, research investigations should establish treatment conditions and test their effectiveness in controlled situations. Presently, few researchers have attempted to do this.

Prior investigators (e.g., Colwell, 1995; Rosenthal, 1985) point toward peer interaction and listening to recorded performances as potentially effective for enhancing undergraduates' ability to self-evaluate. We designed the following preliminary investigations to explore the effectiveness of small-group peer interaction combined with sharing of peer feedback on students' ability to self-evaluate their recorded performances. Specifically, we addressed the following research questions:

1. What is the effect of small-group peer interaction combined with peer feedback on undergraduate students' ability to self-evaluate performance accurately?

2. How are correlations among instructor, peer, and self-evaluations affected by small-group peer interaction combined with peer feedback?

EXPERIMENT 1

Method

For the first experiment, we asked faculty area coordinators of brass, strings, voice, and woodwinds at a large Midwestern university to submit a list of all undergraduate music education and performance majors. Thirty participants were randomly selected from all undergraduate music education majors within these performance

areas, and all agreed to participate. (Because of difficulties with distance and location, it was not feasible to include percussionists in these studies.) The number of available music education majors in the string and voice areas was small; therefore, undergraduate music performance majors were randomly selected to complete the groups. One withdrew, leaving 29 performers. These included 6 brass instrumentalists; 6 string instrumentalists, of whom one was a performance major; 6 vocalists, of whom 2 were performance majors; and 11 woodwind instrumentalists. Fifteen of these students (3 brass, 4 string, 3 vocal, and 5 woodwind) were selected randomly for the experimental group. The remaining fourteen made up the control group.

With approximately 3 weeks left in the semester, we assigned the experimental group into four groups of three to five students. Assignment was based on performing medium (i.e., one group consisted of vocalists, and another, of brass instrumentalists, etc.). During a 2-week period, each small group met in four separate sessions, during which students played or sang a solo piece of their choosing. Most students performed music they were working on for end-of-semester performance juries. Each session was videotaped with high-quality videocassette recording equipment, with recording order for each session randomly determined. Immediately following the video recording, students viewed the videotapes collectively, briefly discussed the performances, and completed peer and self-evaluations. Peer evaluations were shared with participants after self-evaluations were completed. Peer- and self-evaluation forms were similar to the forms used in Bergee's (1997) study. Categories included those regularly established in facet-factorial studies of music performance structure (e.g., Abeles, 1971; Bergee, 1988; Jones, 1986): tone, intonation, technique (with articulation, bowing, and diction specified as appropriate), and interpretation/musical effect. Students were instructed to score each category from 0 (poor) to 10 (excellent). At the bottom of the form, space was provided for students to write comments.

To avoid duplicating the instructor-centered private lesson environment that Hepler (1987) has described, facilitation was kept to a minimum. The facilitator only provided structure and helped participants remain focused on the task at hand. Researchers studying peer evaluation of teaching (e.g., Razelle, 1998) have commonly allowed peers freedom to discuss teaching episodes openly and to arrive at their own decisions regarding quality, effectiveness, and so forth.

The four sessions concluded the week prior to the end-of-semester performance juries. Within the same time frame, control-group students participated in no special activities. We used participants' jury performances for evaluation, which were videotaped. Immediately following the performance jury, participants were escorted to a second location, at which they viewed the videotape and completed self-evaluations. In addition, students in the experimental group were asked to comment anecdotally about the small-group sessions.

Faculty members and graduate students evaluating the jury performances were those normally responsible for end-of-semester evaluations ($n_{brass} = 5$, $n_{string} = 4$, $n_{voice} = 6$, and $n_{woodwind} = 4$). In addition to the normal jury evaluation, faculty members completed the four-category evaluation form described above. Interjudge reliability among the faculty panels was acceptable, ranging from .87 to .93 (*W*s, *p*s < .05). The unit of analysis was deviation of students' self-evaluation scores from averaged faculty scores. Correlation coefficients were calculated between individuals' self-evaluation and averaged faculty evaluation.

Results

Given that we were exploring the effect of an untested treatment rather than confirming or corroborating an existing one, we set the region of rejection (alpha) at .10. Preliminary analyses showed no significant differences in ability to self-evaluate by year in school, performance level, or medium. Multivariate analyses with tone, intonation, technique, and interpretation/musical effect serving as dependent variables resulted in an overall significant difference, $F(1, 28) = 2.21$, $p = .09$; *lambda* = .72. Follow-up univariate analyses revealed a significant difference in intonation ($F = 5.34$, $p < .03$) and no significant differences in the other categories (tone $F = 1.43$, $p = .24$; technique $F = 2.55$, $p = .12$; interpretation/musical effect $F = 0.04$, $p = .85$). There was considerable variability of self-evaluation deviation within both groups on all performance criteria, especially on tone and interpretation/musical effect.

Correlations between self-evaluation and others' evaluation suggested that the experimental group was able to self-evaluate somewhat more accurately than the control group, with the exception of tone. Ability to self-evaluate technique seemed the most secure in both groups. Faculty-self correlations in the experimental group were .36, .77, .82, .57, and .69 for tone, intonation, technique, interpretation/musical effect, and total score respectively; for the control group .46, .59, .72, .52, and .58 respectively. (Correlations beyond .31 are statistically significant, $p < .10$.)

Discussion and Critique

Consistent with prior research (Colwell, 1985; Rosenthal, 1985), peer interaction combined with feedback showed initial promise in improving participants' ability to self-evaluate. Not surprising, the ostensibly more "subjective" aspects of tone quality and interpretation/musical effect seemed the least amenable to influence. Correlations corroborated this: self-evaluation of technique, arguably the performance aspect least influenced by individual opinion, was most closely aligned with faculty evaluation.

For the next study, however, the four performance categories (tone, intonation, technique, interpretation/musical effect) re-

quired more definition. Students clearly had divergent conceptualizations, especially of interpretation/musical effect. For this second study, we added more descriptive information to the evaluation form and introduced exemplary recordings to serve as assessment anchors.

We also reconsidered the recording and playback medium. A number of students commented that videotape playback did not accurately portray their performing (see Daniels, 2001, several of whose respondents to a performance self-assessment survey commented similarly). For the second study, we used aural playback only, recording students' performances directly onto digital compact discs.

During the course of Experiment 1, it became apparent that assessing changes in self-evaluation ability over time rather than focusing on one summative evaluation would provide a truer picture of changes in self-evaluation ability. Furthermore, the stress involved in jury performances may have distorted self-evaluations. Therefore, in the following study we compared self-evaluation to others' evaluation across multiple sessions. We asked students to peer and self-evaluate in more informal and less stressful settings, and we tied evaluations directly to the music performed in these settings.

Finally, a strong need existed to address the high within-group variability. Such extreme variability may have masked the effectiveness of the treatment condition. Following Kerlinger's (1986) principle of maximizing systematic variance and minimizing error variance, we added an independent variable: extent of self-evaluation deviation from others' evaluation. For Experiment 2, we blocked participants into high and low self-evaluation deviation (with the first self-evaluation session's median deviation scores serving as the cutoff), randomly assigned participants to two groups, and randomly assigned one of the groups to the treatment condition. Therefore, Experiment 2's was a mixed Treatment by Blocks design consisting of two between variables (group, self-evaluation deviation) and one within (time, i.e., self-evaluation across five sessions).

EXPERIMENT 2

Method

For the second study, we randomly selected participants from all undergraduate instrumental and vocal music and music education majors enrolled for studio instruction, excluding Experiment 1 participants. Of the 80 contacted, 71 agreed to participate; 2 of these, however, withdrew from the university before the beginning of the academic semester during which the study was conducted. Of the 69 who began, 13 withdrew, all before the second session. Of the 56 who completed the project, 22 were vocalists, 11 were string instrumentalists, 10 were woodwind instrumentalists, and 13 were brass instrumentalists. Of the 13 who withdrew, 6 were vocalists, 5 were stringed instrumentalists, 1 was a woodwind instrumentalist, and 1 was a brass

instrumentalist. Mortality from the experimental group was 6 and from the control group 7, ultimately resulting in equal groups (*n*s = 28). As in the previous study, we assigned experimental group participants to small groups (3–5) based on their performing medium.

In this study, we used the same evaluation categories as we did in Experiment 1, as well as the same 1–10 rating scale. In an effort to clarify the evaluation criteria, we asked performing faculty and doctoral students to suggest additional descriptors for each of the four broad categories. There was a great deal of overlap between suggested descriptions; we were able to add all descriptions in parentheses under each category. For tone, descriptors included depth, richness, characteristic sound, and consistency throughout range. For intonation, descriptors included consistency within scale and phrases, and accurate interval relationships. For technique, descriptors included tonguing, bowing, diction, fluency of finger movement, breath support, placement, vowel selection, diction, and accurate rhythm. (We pointed out to participants that not all descriptors applied to every performing medium.) Under interpretation, descriptors included expressiveness, adherence to articulation markings, adherence to style, use of dynamics both written and implied, use of melodic shaping, appropriate use of vibrato, and phrasing.

To help establish a more consistent frame of reference for performance evaluation, we asked faculty to suggest recordings they considered exemplary in all four categories, especially interpretation/musical effect. Of the recordings suggested, we chose two, one vocal and one instrumental, that several faculty had strongly recommended: Anne Sofie von Otter's recording of Brahms's "Ach, wende diesen Blick," Op. 57 no. 4 (DG 429 727-2 GH), and Nathan Milstein's recording of J. S. Bach's *Partita No. 2 in D* for solo violin, BWV 1004 (DG 423 294- 2 GCM2).

All participants self-evaluated a total of five times. During Session 1, actually a pre-session, participants individually played or sang all or part of a solo piece (unaccompanied) of their choosing. Performances in all sessions were recorded on a Sharp MD-MT15 minidisc portable digital recorder. Immediately following the recording, participants listened to the playback and completed a self-evaluation form.

The remaining sessions took place about weekly. For Sessions 2 to 5, experimental-group participants met in their small groups and listened to excerpts from the two "anchor" recordings. After discussion of the performing qualities displayed on these recordings, participants played or sang all or part of a solo piece of their choosing, which was recorded. From this point, procedures were the same as Experiment 1's.

Participants in the control group met individually with the facilitator during time frames corresponding to experimental-group sessions. They sang or played all or part of a solo piece, which was recorded on the minidisc recorder. Immediately following this, participants listened to the recordings and completed a self-evaluation.

The two experimenters served as evaluators (hence "experimenter evaluation"). One of the two experimenters acted as the facilitator; the other, however, was largely unfamiliar with the performers. Total score interrater reliability (r) was .91 ($p < .01$), and category reliabilities likewise were acceptable (.85 to .94, $ps < .01$). As an independent reliability check, a panel of graduate students in music education ($N = 8$) evaluated a randomly selected 10% of the performances. Intrapanel reliability was good for each of the categories, ranging from .86 for intonation to .93 for technique (Ws, $ps < .01$). Total score intrapanel reliability was .90 ($p < .01$). Category interjudge reliability between panel and experimenter evaluation ranged from .87 to .95 (rs, $ps < .01$); total score reliability was .92 (r, $p < .01$).

Results

Owing to the extensiveness of the statistical testing, we chose .01 as the alpha level for Experiment 2 analyses. Preliminary analyses showed no significant differences between experimental and control groups on Session 1 self-evaluation deviation. After determining that the assumptions of multivariate normality and homogeneity were met, we conducted multivariate analyses on the data with the four categories of tone, intonation, technique, and interpretation/musical effect serving as dependent variables.

Between-subjects main effects of group (experimental/control) and self-evaluation deviation (SED) were not statistically significant ($F = 1.91$, $p = .12$; $F = 3.02$, $p = .03$ respectively), nor was the interaction between group and SED ($F = 0.18$, $p = .95$). The within-subjects main effect of time (i.e., evaluation across the five sessions) was not significant ($F = 1.30$, $p = .25$), nor were the time by group interaction ($F = 0.42$, $p = .97$) and the time by group by SED interaction ($F = 0.58$, $p = .88$). The time by SED interaction, however, was statistically significant ($F = 3.73$, $p < .001$, *lambda* = .37). Univariate tests on the time by SED interaction showed significant differences on tone ($F = 5.60$, $p = .001$, $eta^2 = .10$), intonation ($F = 5.46$, $p < .001$, $eta^2 = .10$), and technique ($F = 3.55$, $p = .008$, $eta^2 = .06$), but not on interpretation/musical effect ($F = 0.89$, $p = .47$). A graph of the significant time by SED interaction for technique is located in Figure 1 (tone and intonation contours were similar). SED declined substantially at the second session for high SED participants and gradually increased afterward for both high and low SED participants.

With the exception of interpretation/musical effect, experimental group/high self-evaluation deviation participants' self-evaluation deviation tended to decrease across the sessions (e.g., tone from $M = 2.19$, $SD = 1.03$ for Session 1 to $M = 1.16$, $SD = 1.30$ for Session 5; intonation and technique were similar). Control group/high self-evaluation deviation participants showed a similar but less pronounced tendency (e.g., tone from $M = 1.50$, $SD = 0.84$ for Session 1 to $M = 1.36$, $SD = 1.05$ for Session 5; intonation and technique were similar). In both experimental and control groups, low self-evaluation deviation

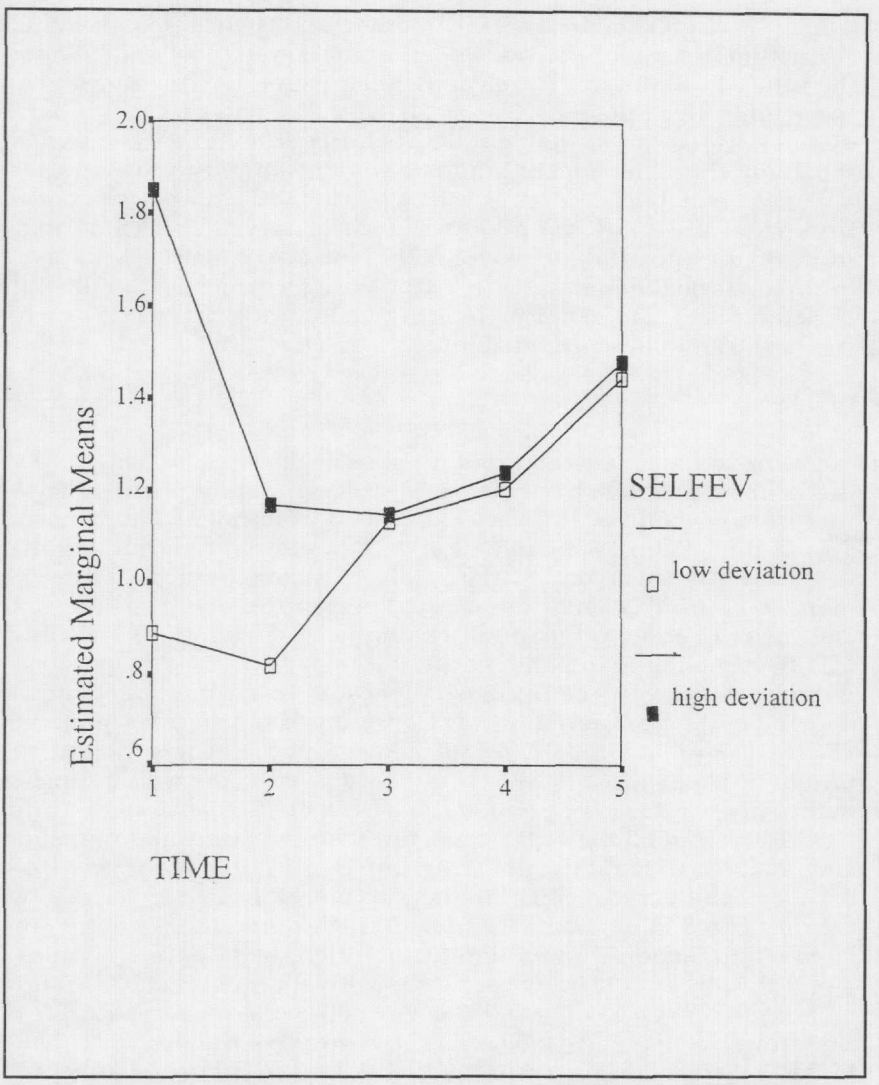

Figure 1. Graph of the significant time by self-evaluation deviation interaction for technique.

participants showed a consistent increase in self-evaluation deviation across the five sessions (e.g., combined tone from $M = 0.74$, $SD = 0.73$ for Session 1 to $M = 1.35$, $SD = 1.15$ for Session 5). In terms of raw scores, experimenter evaluation and peer evaluation showed no pattern of increase or decrease over time. Peer evaluation, however, was consistently higher than experimenter evaluation. Over the five sessions, self-evaluation scores increased, moving away from experimenter evaluation and toward peer evaluation.

Correlations among experimenter and peer evaluations initially were statistically significant and positive for the total score and for three of the four categorical scores. By Session 5, however, there were no statistically significant correlations among experimenter and peer evaluations. Correlations between peer and self-evaluations initially were low but by Session 5 had attained statistical significance on total and all categorical scores. Correlations between experimenter and self-evaluation generally were low, showed no pattern of increase or decrease over time, and showed no marked differences from correlations between experimenter and self-evaluation in the control group. Interrater reliability within the small peer groups was uneven, ranging from .13 to .88 (*W*s).

Discussion

A treatment consisting of small-group interaction combined with peer feedback seemed not to have had a strong effect on self-evaluation skills. A number of factors may have accounted for this. The modest observed power for the group variable (.64) suggests that a larger sample might have helped. Perhaps a time frame extended over an entire semester or academic year would have led to greater differences. Most likely, however, the treatment was not a powerful enough intervention to effect better self-evaluation. We avoided a strong instructor/experimenter presence, and for similar reasons we avoided direct sharing of instructor or experimenter feedback with students. We were concerned that students would adjust self-evaluations to match instructors', regardless of whether the students had truly heard and internalized the discrepancies.

Such nondirectiveness should be reconsidered, given the apparent difficulty of accurate self-evaluation and the novelty of asking students to assess their own and their peers' performances. Future experimenters should consider more direct instructor/experimenter oversight of the self-evaluation process, perhaps with strong initial instructor/experimenter involvement and then regular follow-up sessions to determine whether enhancements in ability to self-evaluate had sustained themselves over time. Such careful scaffolding often is recommended in the early stages of a learning process (e.g., Brown, 1999).

As contrasted with listening to themselves on videotape as in Experiment 1 (which led to speculations as to the quality of the sound and perhaps confounded listening with watching), participants in the second experiment seemed satisfied that the high-quality digital media accurately captured their sounds. Thus, listening on high-quality digital media may have a strong and immediate positive impact on performance self-evaluation ability (see Figure 1). This conclusion is provisional, however, as there was no control group that self-evaluated without listening to recordings. The attention paid to self-evaluation per se might explain this outcome. Future investigators should consider examining the effect of listening to recorded

performances on ability to self-evaluate through extension of the experimental/control approach used in this study.

The initially strong and positive effect on self-evaluation of listening to recorded performances seemed to dissipate over time (see Figure 1). Peer evaluation shared with performers may have adversely affected ability to self-evaluate. In the second experiment, peer evaluation was consistently higher than experimenter evaluation, a phenomenon regularly noted in studies of teaching (e.g., Byo, 1990; Colwell, 1995) and performance (Bergee, 1993, 1997). Peer evaluation shared with performers may lead to inflated and unrealistic perceptions of performance achievement, the condition against which Atwater and Yammarino (1996) have cautioned. This study's participants seemed readily persuaded that relatively high peer evaluations accurately reflected their performing. Beyond its effect on self-evaluations, the sharing of peer evaluations with performers seems to have affected its (peer evaluation's) internal consistency and weakened its relationship with experimenter evaluation. Therefore, it should be used judiciously. We recommend that peer interaction and evaluation be implemented when feasible as a component of such supportive peer-learning activities as peer tutoring (e.g., Arrega-Mayer, 1998), but we also recommend that numerical evaluations not be used, or at least not shared.

As with the previous two studies (Bergee, 1993, 1997), the relationship between others' evaluation and self-evaluation was not strong. Lack of ability to self-evaluate seems persistent and not readily ameliorated. Improvement will require consideration of a broader context. Self-evaluation likely involves a complex interplay of influences, some attributional and others environmental. Atwater and Yammarino (1996) have developed a self-assessment model placing proposed influences under broad categories of biographical characteristics, individual and personality characteristics, cognitive processes, context, and job-relevant experiences. These authors provided substantial research support for their model, which might serve as a starting point for attempting to portray the complexities of performance self-assessment. We recommend that Atwater and Yammarino's model be examined for its utility in explaining variation in ability to self-evaluate performance.

To minimize the effect of extraneous variables, we chose participants to be as homogeneous as possible. This, by definition, limited generalizability. Subsequent studies should broaden participation to include other groups of performers. These studies' provisional findings would gain substantiation if replicated among performers of different levels of maturity and experience.

REFERENCES

Abeles, H. F. (1971). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education, 21*, 246–255.

Arnold, J. A. (1995). Effects of competency-based methods of instruction and self-observation on ensemble directors' use of sequential patterns. *Journal of Research in Music Education, 43,* 127–138.

Arrega-Mayer, C. (1998). Increasing active student responding and improving academic performance through classwide peer tutoring. *Intervention in School & Clinic, 34,* 89–94.

Atwater, L. E., & Yammarino, F. J. (1997). Self-other rating agreement: A review and model. *Research in Personnel and Human Resources Management, 15,* 121–174.

Balch, W. R. (1998). Practice versus review exams and final performance. *Teaching of Psychology, 25,* 181–185.

Benson, W. L. (1989). The effect of models, self-observation, and evaluation on the modification of specified teaching behaviors of an applied music teacher. *Update: Applications of Research in Music Education, 7* (2), 28–31.

Bergee, M. J. (1988). Use of an objectively constructed rating scale for the evaluation of brass juries: A criterion-related study. *Missouri Journal of Research in Music Education, 5* (5), 6–25.

Bergee, M. J. (1993). A comparison of faculty, peer, and self-evaluation of applied brass jury performances. *Journal of Research in Music Education, 41,* 19–27.

Bergee, M. J. (1997). Relationships among faculty, peer, and self-evaluations of applied performances. *Journal of Research in Music Education, 45,* 601–612.

Brown, K. J. (1999). What kind of text—for whom and when? Textual scaffolding for beginning readers. *The Reading Teacher, 4,* 292–307.

Byo, J. L. (1990). Recognition of intensity contrasts in gestures of beginning conductors. *Journal of Research in Music Education, 38,* 157–163.

Cassidy, J. W. (1993). A comparison between students' self-observation and instructor observation of teacher intensity behaviors. *Bulletin of the Council for Research in Music Education,* no. 115, 15–29.

Colwell, C. (1995). Effect of teaching setting and self-evaluation on teacher intensity behaviors. *Journal of Research in Music Education, 43,* 6–21.

Corder, D. L. (1979). Intermediate and advanced level group instruction in undergraduate applied music: A survey and analysis. *Dissertation Abstracts International, 39,* 4108A.

Daniels, R. (2001). Self-assessment in performance. *British Journal of Music Education, 18,* 215–226.

Falchikov, N., & Boud, D. (1989). Student self-assessment in higher education: A meta-analysis. *Review of Educational Research, 59,* 395–430.

Fitzgerald, J. T., Gruppen, L. D., & White, C. B. (2000). The influence of task formats on the accuracy of medical students' self-assessments. *Academic Medicine, 75,* 737–741.

Hepler, L. E. (1987). The measurement of teacher/student interaction in private music lessons, and its relationship to teacher field dependence/field independence. *Dissertation Abstracts International, 47,* 2939A.

Jones, H., Jr. (1986). An application of the facet-factorial approach to scale construction in the development of a rating scale for high school vocal solo performance. *Dissertation Abstracts International, 47,* 1230A.

Kerlinger, F. (1986). *Foundations of behavioral research.* New York: Holt, Rinehart and Winston.

Randall, R., Ferguson, E., & Patterson, F. (2000). Self-assessment accuracy and assessment center decisions. *Journal of Occupational & Organizational Psychology, 73,* 443–459.

Razelle, F. (1998). Suspending disbelief: Improving writing, research, and team-building skills in a peer-centered learning environment. *Journal of Management Education, 22,* 368–386.

Rosenthal, R. K. (1985). Improving teacher effectiveness through self-assessment: A case study. *Update: Applications of Research in Music Education, 3* (2), 17–21.

Ross, S. (1998). Self-assessment in second language testing: A meta-analysis and analysis of experiential factors. *Language Testing, 15,* 1–20.

Walker, M., & Warhurst, C. (2000). "In most classes you sit around very quietly at a table and get lectured at ...": Debates, assessment, and student learning. *Teaching in Higher Education, 5,* 33–49.

Yarbrough, C. (1987). The relationship of behavioral self-assessment to the achievement of basic conducting skills. *Journal of Research in Music Education, 35,* 183–189.

Submitted October 29, 2001; accepted May 28, 2002.