MEASUREMENT AND EVALUATION IN MUSIC EDUCATION: A POSITION PAPER

Diana Santiago[1]

INTRODUCTION

Educators, administrators and parents alike agree that evaluation of instruction is an important feature of the educational process. For the educators, evaluation is a tool to help them improve instruction; for administrators, parents and taxpayers, it is the instrument for appraising the quality and worth of programs, in order to avoid waste of time and/or money. There is not a general consensus, however, on the appropriate procedures for evaluating. This is particularly true in music and in the arts in general, where the affective domain assumes an important role in the curriculum. Tests, the traditional instrument of measurement, are advocated by a large group and condemned by another large group, both of which vehemently present arguments to support their positions.

Evaluation encompasses all aspects of the educational process. It includes student evaluation, teacher evaluation, program evaluation and research evaluation. To make things even more difficult, each one of these subjects can be evaluated at different levels. For example, students can be evaluated individually or in groups, by themselves, by their peers, or by their superiors in the educational hierarchy, and so forth. Add to this picture the fact that evaluation assumes a preponderant position in the performing life of students and artists, and the amplitude of the subject is evident.

Although not claiming to be all-embracing, this paper attempts to outline the major points of tlhe issue, showing the divergent opinions among specialists and concluding with some suggestions. These suggestions will be directed to Bahia State, Brazil, and to the international music education community.

REVIEW OF THE LITERATURE

The material for this review of the literature on measurement and evaluation was mostly compiled from major journals in music education in North America, but includes chapters from standard books on the subject, a position paper from a major educational association, an article from High Fidelity/Musical America and another one from Educational Review. Several interesting publications on particular aspects of measurement and evaluation were listed and examined, but they had to

[1] Professora da Escola de Música da UFBA. Doutoranda em Música. Trabalho escrito em 1988 enquanto cursava o Mestrado na Eastman School of Music como bolsista da CAPES.

be omitted from discussion because they did not state a specific position. Nevertheless, their large number testifies to the great interest in the field.

According to Fowler (1986), evaluation remains the most controversial and uncertain area in education, despite the fact that arts teachers are constantly evaluating their students. This happens because artists and arts educators fear to apply science to the arts, giving preference to feelings and intuition instead of logic and objectivity. After defining measurement as "a way of making a judgement quantitatively" (p. 9) and evaluation as "a procedure for judging worth" (ib.), Fowler stresses the importance of evaluation as a means of obtaining reliable data of what is occurring in the classrooms. Evaluation is delineated as an instrument for demonstrating the credibility of the arts and their importance in the curriculum. The article then examines the topic under four subheadings: Subjectivity vs. Objectivity, Standardized tests, When and How and For what purpose?. The author includes opinions of several educators. Among these educators, there are arguments both pro and contra evaluation: while some think that arts educators ought to know how to defend their positions with the same tools used by teachers of other subject areas (p. 10), the opposite opinion is defended by others (p. 12), who think that we should use other ways to demonstrate the value of the arts. Fowler discusses the fear shared by many educators that objective evaluation in the arts will place the emphasis on the cognitive domain of the curriculum. He also includes divergent positions on the contribution of the arts to the academic achievement of students (p. 12). The author's position is that evaluation is a convincing tool for persuading skeptics of the importance and value of the arts in education.

Duckworth's article (1974) characterizes evaluation in informal education, by contrasting it to evaluation as it happens in the traditional classroom, and by describing teacher's records of children's growth in informal classrooms. The article also calls music educators to make contributions to "informal evaluation" in their field. The author expresses her position on evaluation through quoting parts of a book by Randall Jarell, making it clear that for her formal evaluation is of no good. One infers from the citation that formal evaluation emphasizes cognition and de-emphasizes affective values; and that informal evaluation provides a more personal response to the person being evaluated.

Barber and Hutchcroft's (1985) article adopts the same position as Fowler (1986): evaluation is a tool that helps teachers convince administrators and the community of the importance of the music program. They think that music educators need to make a case for their programs both quantitatively and qualitatively, by means of standardizcd test scores and other objective data such as enrollment figures, etc., and by "narrative comments of support from students, parents, school and community leaders" (p. 8). One of the arguments defended in the article is that "in order to keep music in the public schocls, music educators must make the effort

to become more organized" (p. 9). No one can disagree with the imperative of good organizational skills as a means of attaining better teaching performance in any area. Statements such as the following, however, are faulty and irrelevant to the subject: "The music teacher usually has a larger class load than the regular classroom teacher. To cover the large number of students in other smaller classes probably means that more than one regular classroom teacher must be hired, so eliminating music classes will not save money" (p. 8). The article is also problematic in the following ways: a) the use of the words "aptitude", "achievement", and "talent" in sequence at the top of page nine, makes the reader wonder what the difference between aptitude and talent would be ("...when measuring the student's aptitude, achievement, or talent"); b) the assertion that "supporting evidence that music programs do indeed enhance the learning process in other basic courses" (p. 9) should not be used as an argument in favor of the need of music programs; c) the list of suggested tests is not updated, although the authors refer the readers to the Buros Mental Measurement Yearbook for additional information. The authors express their opinion that tests are useful in discovering musical talent. They also emphasize the importance of matching the evaluation with the music educator's and the school's philosophies and objectives.

Warnick opens his article (1985) by stating that music educators find measurement and evaluation "difficult to relate to the aesthetic of their art" (p. **33),** although he does not include evidence that support the statement. After doing so, he questions why this happens, if musicians are constantly making judgements, and he points to the need for improving the knowledge of future music educators on the subject. This is, indeed, the purpose of the article: not only to point out the deficiency in the undergraduate curriculum but also "to stimulate practicing music educators to become more aware of measurement and evaluation" (p. 34). After contrasting measurement and evaluation, Warnick goes on to explain factors for consideration when selecting a test; thus, he shows the different characteristics of achievement and aptitude tests. The concepts of validity and reliability are explained, including observations concerning the coefficient of correlation. The question of practicability of a test is also discussed in its various aspects such as length, administration, scoring, cost, aural stimuli, instructional clarity, and standardized norms. Next, he explores the measurement of aptitude and of achievement, giving reasons for the necessity of standardized measurements of musical aptitude, and listing the difficulties present in the construction of an achievement test. These difficulties result from the very nature of music and the "lack of consensus among music educators as to what should be taught" (p. 37). He ends the article by reaffirming the importance of the study of measurement and evaluation techniques in the training of music educators, and by saying that such informal evaluations such as observations, inventories, teacher-made tests, and general subjective evaluations should supplement standardized test data. A

descriptive list of recommended achievement and aptitude tests is included. The article is very clear and very appropriate for an overview of the subject at the introductory level. The author firmly believes that an understanding of measurement and evaluation techniques can contribute for an improvement of the music program.

In the second chapter of their book Measurement and evaluation of musical experiences, Boyle and Radocy (1987) analyze contemporary issues in the subject. The chapter provides an overview of the criticisms, misuses and limitations of test, and includes other issues such as the use of norm- and criterion- referenced tests; minimum competency testing; limitations and dangers of testing; extraneous variables influencing test scores; moral, social, and political implications of testing; the global versus specific approaches to music testing; and the behavior versus cognition or affect inference issue. The chapter, like the book as a whole, is very well documented, with an expressive reference list. The discussion on the criticisms of testing includes test quality and misuses of tests. Concerning criterion- and norm-referenced tests, the authors' position is that neither one is superior to the other, both serving different purposes. This position radically differs from Gordon's (1984), who considers that criterion-referenced tests have "severe limitations" and that their use serves to "establish or to perpetuate mediocrity in education" (p. 284). Boyle and Radocy question an overdependence on competency tests for evaluating a school program (p. 35), but nevertheless acknowledge their prominence in current educational practices. They consider tests to be "an essential part of educational evaluation" (p. 38), despite their danger and limitations. In chapter fourteen, the last of the book, the authors speculate about areas that are open for future developments in the field, including competency testing, musical biases related to cultural differences, and technological developments that may affect measurement and evaluation techniques. The list of statements that closes the book summarizes the author's positions, which basically says that tests are not infallible, but can be used to provide "enhanced sources of information for making decisions" (p. 317), if they are carefully constructed and administered.

Perrone's position paper On standardized testing and evaluation (1976), written for the Association for Childhood Education International (ACEI) and for the National Association of Elementary School Principals (NAESP), is the current available position paper of ACEI, which makes us wonder about the permanence of the issues concerning the topic during the last decade. He firmly positions himself against the use of standardized tests. He questions the fairness of such tests; the existence of a "normal curve" that can provide a distribution capable of classifying all children; the usefulness of the information provided by tests; their value in helping young children in their learning; and several other issues. He states that teachers feel pressure to teach to tests; that if tests were not given more attention would be given to integrated learning, there would be a broader range of materials and fewer skill-sheets and workbooks; that teachers would prefer to use

the time devoted to standardized testing for educational activities; and that teachers feel that they can assess children's learning in more appropriate ways than through the use of standardized achievement tests. His other concerns include believing that children are labelled according to the results obtained in the tests, and those "below average" suffer of inferiority complexes derived from the remedial settings in which they are placed due to test results. Also, minorities are unjustly treated in these tests. He thinks that the use of standardized tests limits the curriculum to the areas that will te tested, making the arts, for instance, be neglected. Perrone also believes that the students' individual differences are not considered, because the class has to te prepared for the tests as a group. Considering criterion-referenced tests, Perrone thinks they are an improvement if compared with norm-referenced, but they have significant limitations, which he lists (p. 12). This position disagrees with Boyle and Radocy's and with Gordon's positions, as stated in the previous review. Perrone's position is that "a moratorium is needed" (p. 12) for all standardized tests, so that a development of alternative forms of evaluation can take place, as well as a critical reexamination of the various issues concerning tests. Perrone considers evaluation basic to educational growth; he wants it, however, to be consonant to what is happening in the classroom. As alternatives to standardized testing, he offers some directions in the last part of the paper. Suggesting a systematic process of documentation as a better option, he questions if standardized achievement tests "reveal as much as carefully kept records maintained over a period of time" (p. 14). The paper is well structured and documented and delivers well what it proposes.

The article by Hargreaves (1974) relates to psychological testing. I consider appropriate to include it here because the criticisms applied to such tests can be applied to aptitude tests, and so it is important to examine them. In this article, Hargreaves examines four main criticisms of psychological tests. The four areas of criticism are the study of intelligence, inherent limitations of I.Q. tests, problems of interpretation and the social implications of testing. Pointing out that a large proportion of this criticism is justified, he nevertheless affirms that tests are the best means for assessing and evaluating individual differences. The "study of intelligence issue" refers to the dichotomy between psychometric theories of intelligence and the study of cognition in general; to the fact that "there is more to intelligence than being able to do well on tests" (p. 28); and to the argument presented by some critics that I.Q. tests measure intelligence according to arbitrary educational standards. The inherent limitations of I.Q. tests consist of problems of cultural bias and the establishment of standardized norms. Interpretation of test scores must take into consideration not only non-psychometric aspects of the test situation which affect the scores but also the fact that scores are "indicators of potential ability, and not a fixed characteristic" (p. 29). Considering the social implications of testing, Hargreaves explains that the test user has the responsibility

to safeguard the test taker from abuse of the results of his scores. Hargreaves ends the article by suggesting directions for further development. His position is that the concept of evaluation needs to be broadened, and that tests should not be neglected, but should be part of a wider conception of assessment. If specialists incorporate more naturalistic forms of assessment, the negative aspects of the testing situation such as control, restriction and anxiety "might eventually disappear" (p. 32).

Bates (1984) proposes "a more comprehensive approach to music evaluation" (p. 8), one which takes into consideration all aspects of the music program. Her article underlines the fact that evaluation is fundamental to the educational process, not only as a measurement of students' achievement, but also of programs' accomplishments. Referring to discrepancies in subjective evaluations found in a number of studies and to the "receptive climate today for improvement in evaluative techniques" (p. 6), she offers a list of comments concerning the area. These comments in general agree with Boyle and Radocy's opinions (1987); with Warnick (1985) when he affirms that several techniques should be employed for evaluation, not only standardized tests; and with Hargreaves (1974) in that the concept of evaluation should be broadened. It is worth mentioning that Battes considers grades as "a potential source of positive stimulation in the music class" (p. 7).

Underlining the subjectivity present in all forms of measurement, Radocy (1986) characterizes methods for quantifying the aspects of the musical experience or the musical behavior which cannot be "counted" — such as evaluating faculty, music programs, or musical performances. The article includes an outline of the evolution of the concept of numbers. The psychophysically-based procedures that he suggests should be used in music as appropriate tools to "quantify the uncountable" are magnitude estimation, paired comparison, and the method of successive intervals. The author thinks that "multifaceted data are very useful in enhancing measurement" (p. 23), and includes qualitative descriptions as another useful tool.

I would like to end this review with two standard books on the subject. Although written twenty and eighteen years ago respectively, they still assume a significant position in the related literature. These texts highlight how several aspects of the matter have remained the same since that time.

In the last chapter of his book, Lehman (1968) discusses the major criticisms to testing, the influence of test content on curriculum, and what he considers to be the trend for the next years. His position is that standardized tests are the best tool for measurement because of their objectivity. Against the arguments of test misuse and invasion of privacy, he affirms that any tool can be misused and that "each person should have some control over the manner in which he presents himself in the vocational marketplace" (p. 85). He also mentions the enormous proportion of testing that occurs in the schools, and the oppositions to it. Concerning curriculum,

Lehman thinks that standardized tests tend to emphasize the areas of the curriculum that are tested, but that this is not a good argument in itself for the implementation or expansion of a music testing program. His previsions for further developments are that "music test are likely to become more musical and less atomistic" (p. 86), and that aptitude tests in specialized fields of music are within the possibilities. His final point is that test scores are fallible, that "aptitude tests may be better at predicting failure than at predicting success" (p. 87), and that although a perfect music test might never be attained, all music educators are responsible for improvements in the field.

Colwell (1970) believes that measurement is "probably the best single avenue to improved musical instruction in the public schools" (p. 46), and that evaluation is a teaching tool, the objective of which is not to exclusively assign grades (p. vii). Evaluation is essential for making intelligent decisions (p. 2), although among the limitations of testing is the impossibility of testing all areas of music learning (p. 46). This later position diverges from current opinions (Radocy, 1986), but the reader ought to consider the developments made in the area during the past decade. Colwell also affirms that "music teachers must give more stress to the content of music, and to student achievement in the content, if they wish music to achieve an equal status with English or biology" (p. 47). For this reason, there is a need for "a sound testing program" (ib.). This opinion is also presented by Fowler (1986), who includes its counterpart, as mentioned in the review of his article.

EFFECTS OF THE ISSUE

The large number of standardized music tests written and published in the United States of America demonstrates the interest in the subject, although the review of the literature shows there is little agreement on it. The necessity of a structured evaluation process is stressed by the different sectors of the educational community. Specialists in open education usually condemn standardized tests (Perrone, 1977), but they feel that evaluation is important, particularly on an individual basis (Some commonly asked questions about open education, 1972). Non-traditional measures are employed and developed. Unobtrusive measures — those which do not make use of tests, but collect data from records, physical traces of activities, and observations — are viewed by these specialists and by early childhood specialists as an important tool, one that might decrease the negative effects of testing and enhance the measurement process (Goodwin, W. L., & Goodwin, L.D., 1982).

As mentioned in the introduction to this paper, evaluation embraces all the components of the educational process. As a diagnostic tool, it can reveal very much of what is happening in the classrooms, in the school and in the community as a whole. In order to describe its amplitude, and thus show its effects at the various levels, the present discussion is four-fold. It will include student evaluation,

teacher evaluation, program evaluation and research evaluation. One should keep in mind that by assuming evaluation to be important and necessary, our next step is to take a position on the kind of instruments that should be used in the process. Music educators, individually and as a professional group, must take a position. We cannot rely on the idea that things are going to improve by themselves: evaluation is vital. Next, we have to decide which measurement tools we will use. The process will then be launched, and improvements or rethinkings will take place, from which all society will benefit.

Student evaluation can be done individually or in groups. The measurement in this area includes the following categories: aptitude tests; achievement tests; and attitudes tests, which can be divided into interest tests and preference tests. The music educator — individually and as a profession — must decide: if aptitude tests are valid; if standardized tests are to be used; if it is possible to measure attitudes. The decisions made will exert strong influences on the curriculum.

Teacher evaluation can be done by the teacher (self-evaluation, which also applies to the students, who can self-evaluate themselves); by the students; by specialists in the field, namely other teachers, administrators, supervisors, and researchers. Music educators as individuals must face the importance of self-evaluation and the importance of being evaluated by their students. Then, they have to decide how they want their students to evaluate them. As a profession, music educators must debate the value of teacher testing, and must come to a consensus of how it should be done. I direct the reader to four interesting articles (Culbertson, Schumm, Harrison & Hoit, 1986; Lee, 1985; Michalski, Jr., 1983; Platt, 1986).

Program evaluation includes curriculum evaluation and the evaluation of the efficacy of programs at different levels. Assessment studies are usually developed in order to find out the status quo of such programs. The results of such studies may be applied for diagnostic functions. This is an issue that is currently receiving much consideration due to the stress on the need for accountability in all areas of education. Program evaluation usually receives the attention of educational institutions, in its organization and implementation, but music educators are affected by program evaluation: each teacher will be made responsible for what is found in his or her classroom.

Research evaluation is also relevant to the educational process. It is important that research efforts are optimized, so that the most can be achieved in the least time. It is also important that theoretical research in education be not dissociated from the classroom reality. Music educators can and must assume an important role in the development of research. By trying to keep updated with research findings in their areas, music educators will be able not only to improve their techniques but also to suggest new areas to be studied or to be re-analyzed.

Next, I would like to consider the effects that a pro or contra position towards measurement and evaluation would have. If evaluation is condemned as a whole,

progress in education will be done very slowly. However, if we take for granted the statement made by the Music Educators National Conference that "evaluation is an important aspect of the educational process" (MENC, 1986, p. 55), what really matters is not which side the music educator takes concerning specific kinds of measurements. The most that the position will imply is that different measures will be used, probably because of a different curricular approach. Evaluation will be accomplished, nevertheless, and that is the key factor. Taking a position, and knowing why, is what makes the difference.

POSITION AND SUGGESTIONS

I view evaluation as an essential part of the educational process. Standardized tests are reliable instruments for collecting data, but not the only ones. Unobtrusive measurements must be an integral part of evaluation. I agree with Lehman (1968, p. 84) when he writes that standardized tests are good tools because of their objectivity; I agree with the specialists that stress the need for alternative instruments of measurements because I do not consider standardized tests to be infallible; I agree with Goodwin, and Goodwin (1982) in that unobtrusive measures should not be exclusively used because of "their subtle complexity and their unknown psychometric quality" (p. 540). Thus, I assume a position which is coherent with the assertion that the more measures we employ, the more reliable results we will obtain. This position is in agreement with Bates (1984, p. 8: "comprehensive approach to music evaluation"), Hargreaves (1974), and Radocy (1986, p. 23: "multifaceted data"). My position also reflects my belief that individual differences must be considered, must indeed assume a relevant role in human relationships; this role is of vital importance in the educational process. Tests are not to be used as an instrument of pressure upon the students. Their misuse, however, is not a good argument for their condemnation: "any tool may be misused" (Lehman, 1968, p. 84).

I disagree with educators who think that objective evaluation in the arts will place the emphasis on the cognitive domain (as cited in Fowler's article, 1986): rating scales are an example of objective measurement of aspects of the psychomotor domain, when they are used to evaluate performance. I disagree with Colwell (1970) when he says that "music teachers must give more stress to the content of music, and to student achievement in the content, if they wish music to achieve equal status with English or biology" (p. 47) — this is what would emphasize the cognitive aspect of the curriculum. I do not think that the basic purpose of evaluation is to demonstrate the importance of the arts in the curriculum, as stressed by Fowler (1986) and Barber and Hutchcroft (1985): I believe that evaluation is a tool for improving the quality of music instruction, as Colwell (1970, p. 46) and Warnick (1985, p. 34) also believe.

Teaching is a great responsihility, especially when childhood is involved. Research underlines the importance of the first years of life in the acquisition of the skills and of the dispositions that will command the students' lives (Gordon, 1987; Katz, 1986). Curricular decisions have an enormous impact on students. Given that evaluation plays a significant part on these decisions, its processes must be carefully considered. Educators will benefit more from analyzing the positive and negative aspects of the tools they have to work with, than by condemning them a priori. Perfection is an ideal to be pursued; limitations are a challenge to be conquered.

The following is a list of suggestions concerning measurement and evaluation. The first two suggestions are directed to Bahia State, Brazil. The other two are directed to the international music education community.

- Measurement and evaluation techniques should be discussed in method classes at the undergraduate level and stressed in graduate school;

- A Status quo study of music education in the state should be conducted by the "Associação dos Professores de Educação Musical da Bahia" (Bahia State Music Educators Association);

- An international assessment of music achievement in different countries should be carried out. Besides the potential of this project for stimulating curricular developments, it would increase the number of cross-cultural studies in music education, which is insignificant;

- Research in the measurement of the affective and the psychomotor domains of music should be emphasized.

REFERENCES

Barber, D. M., & Hurtchcroft, C. R. (1985). Evaluation: A positive force. The School Musician, 57(3), 8-9.

Bates, D. (1984). An approach to evaluation in music. Canadian Music Educator, 26(1), 5-8.

Boyle, J. D., & Radocy, R. E. (1987). Measurement and evaluation of musical experiences. New York: Schirmer Books.

Colwell, R. (1970). The evaluation of music teaching and learning. Englewood Cliffs, NJ: Prentice-Hall.

Culbertson, A. E.; Schumm, C. A.; Harrison, C.; Holt, D. M. (1986). Point of view: Teacher competency tests. Music Educators Journal, 72(9), 31-33.

Duckworth, E. (1974). The "bat-poet" knows: Evaluation in informal education. Music Educators Journal, 60(8), 70-72.

Fowler, C. B. (1986). Evaluation: Pros & cons. High Fidelity/ Musical America, 36(November), 9-12.

Goodwin, W. L., & Goodwin, L. D. (1982). Measuring young children. In B. Spodek (Ed.), Handbook of research in early childhood education (pp. 523-563). New York: The Free Press.

Gordon, E. E. (1984). Learning seguences in music: Skill, content and patterns. Chicago, IL: G.I.A.

Gordon, E. E. (1987). The nature, description, measurement and evaluation of music aptitudes. Chicago, IL: G.I.A.

Hargreaves, D. J. (1974). Psychological testing: current perspectives and future developments. Educational Review, 27, 26-33.

Katz, L. G. (1986). Current perspectives on child development. Council for Research in Music Education, 86, 1-9.

Lee, R. T. (1985). Would your students give you a passing grade? Music Educators Journal, 71(9), 60-64.

Lehman, P. R. (1968). Tests and measurements in music. Englewood Cliffs, NJ: Prentice-Hall.

Music Educators National Conference. (1986). The school music program: Descriptions and standards (2nd ed.). Reston, VA: Author.

Michalski, S. F., Jr. (1983). The best you can be: Criteria for self-evaluation. Music Educators Journal, 70(1), 58-59.

Perrone, V. (1976). On standardized testing and evaluation: An ACEI/NAESP position paper. Washington, DC: Association for Childhood Education International.

Perrone, V. (1977). The abuses of standardized testing. Bloomington, IN: The Phi Delta Kappa Educational Foundation.

Platt, M. C. (1986). Where will teacher testing lead us? Music Educators Journal, 72(9), 28-30.

Radocy, R. E. (1986). On quantifying the uncountable in musical behavior. Council for Research in Music Education, 88, 22-31.

Some commonly asked questions about open education. (1972). Music Educators Journal, 59(3), 49-50.

Warnick, E. M. (1985). Overcoming measurement and evaluation phobia. Music Educators Journal, 71(8), 32-40.