THE DEVELOPMENT AND TESTING OF AN INSTRUMENT

to evaluate aesthetic judgments

APPROVED:

Graduate Committee:

[Signatures]

Major Professor

Minor Professor

Committee Member

Committee Member

Committee Member

[Signatures]

Chairman of the Department of Art

Dean of the Graduate School
THE DEVELOPMENT AND TESTING OF AN INSTRUMENT 
TO EVALUATE AESTHETIC JUDGMENTS

DISSERTATION

Presented to the Graduate Council of the
North Texas State University in Partial
Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Mary Alice Brumbach, B. A., M. A.
Denton, Texas
December, 1978

This study was concerned with the development and testing of an instrument to measure levels of aesthetic judgment making. The review of related literature demonstrated that the majority of instruments for aesthetic judgment employed a naive to sophisticated judgment comparison to determine levels of aesthetic sensitivity. The inadequacy of a score reporting only the degree of agreement between the subject's choice and the choice of a panel of experts without indicating the source of agreement was discussed. Content analysis of aesthetic responses used in research studies by Wilson and Morris were presented as an alternative means for determining aesthetic criteria.

The review of evaluation methods for aesthetic judgment resulted in a two-part instrument. Part one required the subject to select the better of two art works and to state the reasons for the choice. Part two, a self-scoring component, consisted of the Wilson categories presented as typical statements containing the primary criterion for the category. The subject was instructed to select the statements that were closest in meaning to his initial response.
Three versions of the instrument, which differed only in the composition of the slide pairs, were given to three groups of math students enrolled in a large metropolitan community college district. Each group was given one version of the test twice with a three-week time lapse between administrations. One hundred fifty-five students were present for the first administration and ninety-five students were present for the second testing period. Art appreciation or humanities survey course students were eliminated from this sample.

Comparisons for interscore and interscorer reliability were conducted for all three test versions. Overall results produced by the instruments were compared to determine the degree of equivalence for the instruments. The number and kind of responses placed in the miscellaneous category by the expert scorer were reported for content validity evidence.

Research data were reported as percentages of agreement, the number of exact matches in categories compared to total responses, between results for each set of scores on each of the three versions of the instrument.

It was found that the categories for aesthetic judgments developed by Wilson produced percentages of agreement for expert scoring and re-scoring of the first set of written
statements ranging from ninety to ninety-five percent. The self-scoring component produced very unreliable scores with (1) student inability to select appropriate categories; (2) students choosing categories unrelated to the written statement, and (3) students changing reasons for aesthetic decisions between the two administrations as the probable causes. The three test versions could not be considered equivalent as a result of differences in percentages of agreement ranging from eight to forty-nine percent.

Ninety-seven percent of the initial one hundred fifty-five responses to the instrument contained aesthetic criteria and ninety-two percent were classified by the Wilson categories. One new category related to viewer effort in comprehending the painting was reported.

This report concluded that the instrument possessed considerable content validity, and the categories are highly reliable when scored by an expert scorer. The unreliability of the self-scoring equipment seemed to be a result of the tendency of directed encounters with art works to change the viewer's perceptions. Further attempts to determine the stability of the instrument by other means were recommended.
# TABLE OF CONTENTS

LIST OF TABLES .................................................. v

Chapter

I. INTRODUCTION ................................................. 1

Origin of the Study
Objectives of the Study

II. REVIEW OF RELATED LITERATURE ....................... 12

Introduction
Appreciation and Aesthetic Judgment Making
Instruments for Measuring Aesthetic Judgment
Problems and Possibilities of Aesthetic Judgment Tests
Requirement for Aesthetic Judgment Instruments
Summary

III. DESIGN OF THE STUDY ................................. 42

Introduction
Statement of Purpose
Statement of the Problem
Development of the Instrument
Considerations in Determining Validity and Reliability
Research Design for Reliability and Validity Studies for the Instrument
Hypotheses
Limitations of the Study
Assumptions
Definition of Terms
Description of the Sample
Procedures for Collection of the Data
Procedures for Treatment of the Data
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>IV. FINDINGS AND INTERPRETATION OF THE DATA</td>
<td>66</td>
</tr>
<tr>
<td>Introduction</td>
<td></td>
</tr>
<tr>
<td>Hypothesis One</td>
<td></td>
</tr>
<tr>
<td>Hypothesis Two</td>
<td></td>
</tr>
<tr>
<td>Hypothesis Three</td>
<td></td>
</tr>
<tr>
<td>Hypothesis Four</td>
<td></td>
</tr>
<tr>
<td>Hypothesis Five</td>
<td></td>
</tr>
<tr>
<td>Hypothesis Six</td>
<td></td>
</tr>
<tr>
<td>Findings</td>
<td></td>
</tr>
<tr>
<td>V. SUMMARY, CONCLUSIONS, IMPLICATIONS AND RECOMMENDATIONS</td>
<td>80</td>
</tr>
<tr>
<td>Summary</td>
<td></td>
</tr>
<tr>
<td>Conclusions</td>
<td></td>
</tr>
<tr>
<td>Implications</td>
<td></td>
</tr>
<tr>
<td>Recommendations for Further Research</td>
<td></td>
</tr>
<tr>
<td>APPENDIX A</td>
<td>94</td>
</tr>
<tr>
<td>BIBLIOGRAPHY</td>
<td>98</td>
</tr>
</tbody>
</table>
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>I. Percentages of Agreement for Expert Scoring and Rescoring of Test Versions A, B, and C</td>
<td>67</td>
</tr>
<tr>
<td>II. Percentages of Agreement for Self-Scoring and Expert Scoring for Test Versions A, B, and C on the First and Second Administrations</td>
<td>68</td>
</tr>
<tr>
<td>III. Percentages of Agreement Between Results of the Self-Scoring Component for Test Versions A, B, and C in the First and Second Administrations</td>
<td>70</td>
</tr>
<tr>
<td>IV. Percentage of Agreement Between Results of the Expert Scoring for Test Versions A, B, and C in the First and Second Administrations</td>
<td>74</td>
</tr>
<tr>
<td>V. Amount of Difference Between Percentages of Agreement for the Three Test Versions in Four Scoring Comparisons</td>
<td>75</td>
</tr>
</tbody>
</table>
CHAPTER I

INTRODUCTION

Origin of the Study

A common feature of most two-year and four-year college curricula is a course entitled Art Appreciation. The course is usually "open to all" (9) and has as its stated goal development of appreciation for the visual arts as well as "critical evaluation of selected works" (9). Students register for the course for a wide variety of reasons which may include fulfilling degree requirements, general enrichment, and as a re-entry point for adult learners to academia.

There are some definite knowns in this instructional setting. The course is usually one semester in length, averaging thirty-seven classroom hours. The student population, which may range from eighteen to sixty-five years of age, is frequently drawn from all areas of the college if the course serves as a degree requirement in the humanities for non-art degree programs.

The pattern established in ten semesters of teaching such a course by this investigator and confirmed in the available literature regarding such courses (2, 4) indicates that the majority of students enrolling in Art Appreciation
will share the bias of the general public (10) in believing that "good art looks like the real thing" and that much of modern art is an elaborate hoax. The diversity of the group will, however, result in the presence of numerous other art attitudes. A small number will probably base art values on the moral content of the work. Other participants will determine the worth of an art object on a variety of other bases including the amount of time invested in its production, its uniqueness, or that fact that it is an often reproduced masterpiece (10, 11).

The dilemma presented by this melange of purposes and attitudes is neither a new or an unusual one. Indeed, Edmund Feldman notes,

... today a solution of the problems of aesthetic education of liberal art student still eludes us. The logistics of the situation often impose arrangements which run counter to our educational wisdom. That is, we are governed by the large number of students to be taught, the limited curricular time we have to teach them ... (4, p. 107).

These students have much in common with persons enrolled in similar general education courses. One such group from Macomb College which was described in the ERIC report on curriculum and instruction in the humanities in two-year colleges has the following characteristics:

Previous experience with the Macomb students in other liberal arts courses indicated that they were hostile and skeptical with regard to the fine arts. Although the majority of the approximately 400
students originally selected for the program came from substantial economic backgrounds, only three had ever visited an art gallery or a museum, heard a symphony orchestra perform, or witnessed a professional stage play. In spite of their lack of knowledge or perhaps because of it, the students held tenaciously to negative opinions in regard to the arts and social history in general. They were usually rigid, conservative, and ethnocentric in their judgments (2, p. 92).

There are also some definite unknowns in this teaching situation. Which ones of these students still value realism? Which of these students make art judgments on moral considerations? What other ways of valuing or not valuing art are present in the group? If, as described in the preceding assessment of the Macomb students, there is a general hostility towards art, what are the best ways to deal with that hostility? What teaching strategies would be most effective in such a diverse setting? It is obvious that the realists will require different tactics than will the moralists. Most importantly, given the course goals, how is it possible to determine that they have been reached and to what degree of success in this one semester art appreciation class?

These questions cluster around two primary concerns: (1) course content and strategies for changing aesthetic behavior and (2) evaluation of the effectiveness of those strategies. The first of these concerns has been an area of considerable recent activity in art education research
and writing and constitutes one of the fundamental issues in the field (5). The second concern, evaluation of changes in aesthetic behavior, has received much less attention and, in general, has been hampered by the instruments available to measure such changes (3, 5, 10).

Kathryn Bloom of the John D. Rockefeller III Arts in Education credits a meeting which was funded while she was with the United States Office of Education and directed by Edward Mattil at Pennsylvania State University in 1965 entitled "A Seminar in Art Education for Research and Curriculum Development" as one of the most significant events causing a re-emphasis on aesthetic education (1, p. 94). The seminar brought together experts from psychology, philosophy, sociology, art history and art criticism to join art educators in exploring the basic problems of the field. Five major problem areas were specified at the outset of the ten-day meeting: philosophical, sociological, content, instruction, and curriculum concerns (6). One major outcome of the seminar was a proposal by Laura Chapman, Manuel Barkan and Evan Kern for an aesthetic education project which later developed into the Aesthetic Education Program for the Central Midwestern Regional Educational Laboratory (CEMREL) (1). Both written and oral reports of the developments by the CEMREL staff and discussion and critiques of those products in art education publications and conferences has helped to maintain aesthetic education as a front-runner in art education research.
Ralph Smith in an article entitled "The Philosophical Literature of Aesthetic Education" succinctly states the context of the movement after indicating that it is a revival of an older concern:

Thus, after a period of education in the arts that has relied heavily on the findings and language of psychology, and during which the field has been structured both by images of the child as a creative and performing artist and by expressions that featured process and doing at the expense of the product and subject matter, we are hearing that attention should also be paid to the nature of the product--its properties, meaning, and functions--and to the nature of the appropriate response to art--the character of aesthetic experience, judgment and evaluation (8, p. 138).

The renewed emphasis is also partially a result of a realization that art instruction after the elementary level is all too frequently confined to a "talented group" (1, 7) or to those having academic difficulty. The schism is further fueled by adolescent peer pressure and by frustration with the inability to render well enough to meet expressionistic needs that is typical of this age group (4). The drop in enrollment is basic reality. "Nationally some eighty or ninety percent of the students do not participate in art or music at all while they are in high school" (1, p. 90). This phenomenon appears to many to be a result of curricula structured primarily towards performance and the training of young artists (7). Bloom, along with others, points out graphically,
We know realistically that few of the students graduating from high school will go on to professional courses in the arts. If you consider the much more typical youngster who will spend his life in other ways, what kinds of knowledge and understanding should he have about theatre? music? or the visual arts? . . . let's also remember that while he may become a doctor or a ditchdigger. . . . He may be a member of the schoolboard or even a legislator, and he will inevitably will be a taxpayer (1, p. 99).

Thus, a combined need for audience development, a commitment to a broadened view of art education which includes general education, declining enrollments, and the need to develop curriculum materials for the liberal arts constituency are all factors in the continued importance of aesthetic education. As Hardiman and Zernich report in a summary of art education research from 1970 to 1974, this area of aesthetic behavior is one of the research topics of most activity (5).

However, research efforts in art appreciation related studies antedate the current aesthetic education emphasis by more than twenty years. Davis, in his article on research in art education, summarizes the research concerns from 1949 to 1960 as follows,

The nature of a general factor in aesthetic perception has been of primary concern in research relating to picture preference and aesthetic appreciation since 1940. In the 27 investigations published in the literature between 1940 and 1960, there has also been concern and studies made to determine children's preferences for traditional and modern pictures, as well as the influences of various factors such as prestige and age upon picture preference and appreciation (3, p. 14).
The category of "investigation of picture preference and appreciation" was reported by Davis as one of four major research areas dating from 1883. Hardiman and Zernich emphasize that this research concern has not changed and that investigation of the interaction of variables with picture preference or the effectiveness of varying teaching methods on developing attitudes toward art continue as the prime targets of investigation (5).

The concern for evaluation in art is an often expressed one. As Davis notes, the research area has been one of "interesting activity" (3, p. 14). Prior to 1940, the chief emphasis was on developing standardized tests in drawing, aptitude, and appreciation. In the years since 1940 efforts have been concentrated on the testing of creativity. A few attempts have been made to revise or update tests in appreciation, but this research area "has been virtually devoid of any concentrated or serious efforts" (3, p. 14). Hardiman and Zernich state that "surprisingly few efforts have been undertaken to explain the nature of aesthetic behavior" (5, p. 24). Further, only three of the thirty-three investigations published between 1893 and 1976 have deviated from comparing aesthetic judgments made by art experts to those aesthetic judgments made by non-art trained persons (5, 10). The instruments yield scores which indicate the number of times a naive choice matched expert choices. Unfortunately
the information revealed by these instruments is of little use when confronted with the variety of individuals enrolled in art appreciation courses. Having a score which indicates the degree of sophistication of judgment by a student in this instance is akin to having a winner in a cross-country race without being at all certain of the route taken to the finish line. Knowing the differences between persons basing judgments on moral concerns and those determining art values on the basis of realism requires discerning the route to the decision and the content of that aesthetic decision, not simply that the student made choices which were more or less in line with the experts. This difficulty along with numerous others to be described in the review of the literature on aesthetic judgment instruments is one of the prime motivations for this research effort.

In a summarizing article in the Arts in American Higher Institutions Feldman stated, "It must be plain that the aesthetic education of the entire student body is the most important responsibility of any art faculty of any higher institution. The identification and training of artists and scholars is an important but secondary responsibility" (4, p. 104). Despite this strongly emphasized priority, research and development of aesthetic education curricula, both in content and evaluation, has been almost exclusively confined to the kindergarten through twelfth grade category. Further,
the only major study relating to the art preferences and decision making of adults is an investigation conducted for the United Nations by the Toronto Museum in 1963 (10).

Objectives of the Study

A number of concerns coalesce in this study. First is the renewed emphasis on aesthetic education as opposed to emphasis only on the production of art works by a select few. The lack of concentrated and adequate research efforts in the evaluation of aesthetic behavior is a second problem. Third, the concern expressed by Bloom over a future school-board member or taxpayer is no less cogent to the adult liberal arts student who is probably already a taxpayer and school board member. This probability lends a particular urgency to development of adequate curriculum materials for this segment of the academic population in aesthetic education. Finally, there appears to be no adequate tool for discovering the kinds of art judgments present in such courses as art appreciation or to determine if and how these judgments change. Far too little reliable information is available and thirty-seven hours of class time is a very short period in which to accomplish those changes.

This study will, therefore, be concerned with devising an evaluative instrument to determine aesthetic behavior on a basis other than the comparison of naive to sophisticated judgments. In addition, the instrument will be designed for
adult learners and will be constructed to meet the time constraints imposed by a one semester course.
CHAPTER BIBLIOGRAPHY


CHAPTER II

REVIEW OF RELATED LITERATURE

Introduction

An attempt to develop an instrument which yields useful information regarding the aesthetic decisions of adult learners causes consideration of the following questions in relationship to the literature: What is known regarding appreciation and aesthetic judgment making? What experimental and standardized instruments are available for measuring appreciation? What problems and possibilities exist in these instruments which can be avoided or adapted in the proposed new instrument? What are the requirements to be met in constructing such an instrument? The purpose of this chapter is to present information relevant to these questions.

Appreciation and Aesthetic Judgment Making

The relationship between appreciation and aesthetic judgment making is often a curious and somewhat tenuous one. The term appreciation conveys both awareness and valuing. The object is valued in and of itself and is the subject of intense focus and concentration (28, pp. 160-163). Appreciation also carries a connotation of sensitivity and heightened
awareness as a result of that focus. It follows that the level of appreciation increases with the knowledge of the viewer regarding the thing that is the subject of his attention. The activity does not require a choice. Broudy gives the example of a woman who may appreciate several men without being compelled to choose among them (8, p. 309). Osborne notes that it is possible to treat almost any object, activity, or person in an appreciation mode. It is quite possible to have an acute appreciation of virtuoso skateboarding techniques, a finely tuned engine, or a successful football maneuver without a similar heightened perception of anything labeled art. Those things labeled art according to Osborne are, however, "aesthetic objects specifically designed or particularly adapted to favor prolonged and repeated" aesthetic experiencing (28, p. 161). The difference between the appreciation mode and ordinary perception is perhaps best illustrated in the following statement by Cézanne:

Sometimes I have accompanied a farmer behind his cart driving to market to sell his potatoes. He had never seen Sainte Victoire. They know what has been planted there, along the road, how the weather is going to be tomorrow, whether Sainte Victoire has his cloud cap on or not; they feel like the animals do, like a dog who knows what this piece of bread is only from their needs; but that the trees are green, that this green is a tree, that this is red, and that this red rubble and boulders are hills, I really do not believe that most of them feel that, that they know it outside of their unconscious feeling for the useful (27, p. 10).
Of particular importance is the stress on seeing a place for itself, rather than using it as a source of information.

Aesthetic judgment making requires making a choice or a ranking. The judgment assigns a value to the art work. That assessment of worth is formulated in relationship to the standard of value held by the judgment maker (11, p. 49). Ideally that judgment is based on the investigation of the work itself. Smith and Feldman (29, 16) indicate that the proper sequence of events leading to an informed judgment consists of description, analysis, interpretation, and evaluation. The descriptive phase is concerned with naming, identifying, and classifying. It may include information regarding the kind of object, the materials used, the techniques employed, and relevant historical or iconographical information that identifies the subject matter. Analysis involves a close look at the relationship of parts of the work to each other and to the total work of art. A typical analysis statement might describe the way one color offsets another or the kind of contrasts employed in the painting. The interpretation phase deals with the meaning of the work as a whole, utilizing the results of the analysis as supportive data. Interpretation is never to be confused with narration of the subject matter, but rather embraces both subject matter and the ways in which it is presented. Evaluation is the final step in this pyramid. Like interpretation,
its roots are in the preceding description, analysis, and interpretation phases. The work is deemed successful or unsuccessful on the basis of the value standard employed.

In all facets of this process, the degree of attention to the object is a significant factor. An evaluation based upon full perception of the subtlety and nuances of the work which characterize the appreciation mode will provide a richer and more valid premise for the judgment.

However, things are rarely ideal. Osborne in the *Art of Appreciation* indicates that the making of a judgment regarding an art work is not something necessarily reserved for informed experts:

On the one hand it is taken for granted that appreciation of art and natural beauty is in some sense a matter for everyone and not the prerogative of a restricted class of experts. In the appraisal of the fine arts, as in questions of politics and gastronomy, the man-in-the-street will boldly hazard and affirm his opinion, although in other fields of expertise, such as medicine, or engineering, or the law, he would automatically defer to a professional. On the other hand, society does recognize and support specialists in the arts. They are curators of public galleries and museums, critics, teachers, collectors, dealers, and connoisseurs. Considerable sums of public and private money are spent in the cultivation of the arts. . . . Yet behind the prevailing attitude of ordinary people there is a tacit and true statement that a great deal of knowledge about the arts . . . does not necessarily . . . qualify an expert to impose his aesthetic appraisal on the laymen, and when confronted with a work of art many men will confidently voice their own opinion with the preamble, 'I'm not an expert, but . . .' (25, p. 1).

This lack of hesitation toward the matter of determining the aesthetic worth of an art object is confirmed in the
findings of a study conducted by the Toronto Museum for UNESCO in 1963. The museum personnel, using postcard reproductions of art works, surveyed a random sample of over 500 Toronto residents to determine their preference for art styles and subjects, and the degree of understanding the subjects of this study had of formal elements of art works. Among other findings was "the willingness of the respondents, regardless of their art illiteracy, to state preferences, to make judgments, and especially to say what they do and do not like" (31, p. 134).

From the preceding, it is obvious that numerous aesthetic judgments are made on a less than ideal basis. In actuality a number of premises are used as value standards. A series of categories which summarizes criteria for evaluating art works used by both trained and untrained observers has been developed by Wilson (34) in preparation for a study of the aesthetic behavior of high school students. In addition to devising the categories, Wilson, after a review of the literature in aesthetics, ranked the aesthetic premises on the basis of their adequacy for making aesthetic judgments. Adequate criteria were assumed to be those which were based on the evidence present in the work itself or the situation which surrounded it, while inadequate criteria were those which brought pre-conceived standards such as appropriate moral content or correspondence to reality to the work as
a measuring device. Because these categories offer possibilities for the development of a more informative and precise instrument, they will be presented in detail in the following paragraphs.

Wilson gave the titles **organicism** and **contextualism** to the two most adequate judgmental categories. Organicism is central to the theories of Roger Fry (17) and Clive Bell (6) and is concerned with how well the elements of a work function as a whole. The application of the criterion is basically analytical and involves the relationship of one part to another or one part to the whole work. Examples of statements that relate to the organistic category are: "All the colors blend nicely. Everything fits together. The shapes are interesting. All parts of the work relate to the whole work. Good art work should have a feeling of unity and completeness" (34).

Contextualistic criteria, in contrast, emphasize the total experience of the art object as it affects the viewer. Contextualism was the premise employed by John Dewey in *Art as Experience* (14). Statements in this category include: "The painting is forceful and strong. It has great impact on me. It is very lively. The painting is very expressive of feelings. Good paintings ought to cause an intense and rich experience" (34).
Criteria for judgment that are considered somewhat adequate include hedonism, communication, originality, correspondence to reality, and technical achievement. The designation "somewhat adequate" is based primarily on the work of Stephen Pepper (26). These categories determine aesthetic worth through selected aspects of the work which do not include the whole work of art. The least adequate criteria, moral worth, narrow a priori, authority, and personal preference, generally ignore the art work in the judgment premise.

Hedonism is based on the dimension of pleasure or pain afforded the viewer by the work of art. Work yielding the "greatest amount, the most intense, the longest pleasure" (34) is considered to be the work of highest quality. Its weakness as a criterion stems from the possibility of non-aesthetic concerns such as fond memories or generally warm feelings about the subject matter determining the aesthetic value of the work. Statements classified as hedonistic include: "I enjoy looking at this painting. It gives me a great deal of pleasure. I can imagine that I would enjoy looking at it for quite a while. Good art work should cause feelings or pleasure or pain" (34).

Tolstoy is among those who favored communication as a criterion of aesthetic worth. The following statement from the UNESCO study (31) is indicative of the premise being
employed: "It does not appear that they any longer expect a painting to tell a story or point a moral, but unless it permits an apprehensible image which they can at least feel, they will reject it... They are clearly of the opinion that good art has meaning and is meant to be fairly readily grasped by anyone of average intelligence" (31, p. 134). In its simplest form this criterion might be expressed as "that painter really gets his message across" while a more sophisticated statement might include recognition of the universality of the human experience expressed in the work. This criterion effectively eliminates much of the work of the late 1960's and 1970's as well as any art objects whose concerns are primarily formal (34).

Originality as a category includes the idea of novelty as well as significant departures from established art forms. The criterion's usefulness for evaluating art work is limited by its insistence on uniqueness. The problem is summed up in Beardsley's statement, "It may be original and fine or original and terrible" (4, p. 321). Typical comments for this category include: "It is new and original. It has not been done before. The artist has made something different. Good art should be original" (34).

Craftsmanship is the basis for the category of technical achievement. The criterion is concerned with the skill of the producer as evidenced in the product and not with the product
itself. Typical statements are as follows: "It looks like the artist knew what he was doing. It has taken a lot of time. The painting is done well. Good art work should show that a lot of care has been taken with it."

The classification of aesthetic worth on the basis of correspondence to reality or mimesis severely limits the art objects that can be considered worth of value. As the earlier statement from the UNESCO study affirms, the criterion's use tends to eliminate any work that varies from "actuality" and photographic exactness. This criterion is generally expressed: "It looks like a photograph. It is very realistic. It looks like the real thing. Good art should look like something" (34).

In the least adequate group, the moral worth criterion is concerned with the merit of the subject matter in the painting. In order to be considered good, the subject matter must serve as a vehicle for moral, ethical, social, or political ideals held by the viewer. John Ruskin and William Morris were both staunch supporters of this premise (34). Statements for this category are: "The art work is inspiring. It portrays good ideas. Its message is acceptable. Good art work should be about good things" (34).

The narrow a priori concept includes such statements as "It has everything a painting ought to have. Everything I think a work of art should have is there. My favorite colors
are in the painting. It would go well over my couch. It is my idea of a perfect painting" (34). These indicate that the viewer has already determined a standard for art works that may or may not have any relationship to the work being evaluated. The summary rejection of abstract work reported in the UNESCO study (31) is also a good example of the application of this criterion.

The statement, "Janson lists this work as one of the thousand monuments of art history and it has to be good" belongs in the category of authority. This highly inadequate category can also be expressed in these terms: "Lots of people like this kind of art. It has been shown in magazines and books. Teachers have said it was a good work of art. It looks like an old master painting. It looks like the work they hang in museums" (34). In all cases, the criterion for determining the worth of the object is dependent upon recommendation by some outside authority.

The final category of personal preference can be simply expressed as: "I like it. It appeals to me and I know what I like" (34). Wilson also created a category for his study that was labeled "broom closet" which was used for all statements which did not contain evaluation comments or were irrelevant to the art work (34).

Examination of these categories reveals a direct relationship between considering all aspects of the work, seeing the
art work as a thing in itself, and the most adequate categories of organicism and contextualism. The adequate categories exhibit concentration on some aspects of the work, blocking the full perception of all facets. The least adequate categories essentially ignore the art work and seek a pre-conceived criterion whose existence or non-existence determines the aesthetic value. Child, in a 1964 study titled "Observation on the Meaning of Some Measures of Aesthetic Sensitivity," states that it is possible to infer a high degree of aesthetic appreciation among those persons making valid aesthetic judgments. Conversely, inadequate judgments tend to indicate less aesthetic sensitivity (11, p. 49).

In summary, aesthetic judgments are made on numerous premises, only some of which are related to appreciation or aesthetic sensitivity. Lack of information or art "illiteracy" (31) is no deterrent in making judgments on aesthetic worth, and statements about art works are readily made by non-experts (25). It is apparent that being able to determine the basis for an aesthetic decision, and, therefore, its adequacy does yield information on the level of aesthetic appreciation skills employed by the viewer. Determining the premise for an aesthetic decision could thus be used as measure for success or non-success in meeting the goals of teaching art appreciation.
Instruments for Measuring Aesthetic Judgment

Research relating to picture preference and appreciation has been a consistent emphasis in the literature of art education for more than twenty-five years. In thirty-three investigations published between 1896 and 1976, the major focus has been on "examining the relationship which selected personality variables have to preference or investigating the effectiveness which various teaching methods have on the development and intensity of attitudes" (20, p. 24).

With the exception of two recent studies by Wilson (34) and Morris (23) which have used content analysis techniques, research efforts have employed the results of either a researcher developed instrument or one of the few published tests of aesthetic judgment. The standard procedure in these instruments has been to utilize a comparison of naive to sophisticated judgments to produce the subject's ranking in aesthetic judgment making. In essence, the score a subject achieves translates as the number of times his choices on the instrument agree with those made by a panel of experts.

Most of the research activity related to the development of standardized tests of art appreciation occurred between 1920 and 1940. Among the available tests are the Meier Art Judgment Test (22), the Graves Design Judgment Test (19),
the Beittel Art Acceptance Scale (5), the Child Test of Esthetic Sensitivity (11), the Welsh Figure Preference Test (33), Thorndike's Test for Aesthetic Appreciation (30), and the Christensen and Karwoski Test for Art Appreciation (12).

Each item in the Meier Art Judgment Tests utilizes a pair of reproductions of art objects, one of which is a recognized masterpiece. The other reproduction is the same masterpiece which has been distorted by a variety of means. The subject is instructed to select the reproduction which he considered to be the better of the two art works (22). The Child Test of Esthetic Sensitivity employs reproductions of actual art works. Each pair of examples, such as two Chinese vases, includes one which is considered aesthetically superior (11).

In contrast the Graves Design Test (19) and the Welsh Figure Preference Test (33) use only abstract designs. Each pair of designs has one illustration which has been determined most aesthetically pleasing by a panel of art experts. Both of these were preceded by Thorndike's 1916 Test of Aesthetic Appreciation that utilized rectangles, crosses, and lines within a box frame in groups of five. Each group of five designs was placed in rank order from most pleasing to least pleasing by a panel of judges. The score for the instrument was derived from the number of deviations away from the preferred rank order (30).
The Beittel Art Acceptance Scale (5) consists of a set of art reproductions and a list of statements regarding art works that were classified from naive to sophisticated. This group of statements was the result of responses to the art works collected from college art students (sophisticated) and non-art trained students (naive). Subjects are required to select from this list the statements that best described the art works. Scores on the instrument are determined by the placement of the selected statements on the naive to sophisticated scale.

Beittel's 1953 scale was preceded by a 1926 Test for Aesthetic Appreciation devised by Christensen and Karwoski (12) which employed a variety of illustrations ranging from art objects to architecture and which also listed reasons for choices. Subjects were asked to choose the better of two illustrations as well as to select the reason for the choice. Scores resulted from both correct selection of the more aesthetically pleasing image and the correct reason. Christensen and Karwoski reported that a high correlation existed between the test given with reasons and the test given without reasons. The major difference between Beittel's test and the Christensen-Karwoski instrument is the use of the naive to sophisticated scale by Beittel rather than correct and incorrect responses.
Experimental instruments developed for specific studies are frequently similar to the standardized instruments. Kate Gordon (18) in 1923 used a series of fifty color reproductions of oriental rugs that were placed in rank order of aesthetic merit. Among her findings were that persons ranking the rugs twice were consistent in their choices. However, the panel of experts in this study differed markedly in their assessment of the aesthetic quality of the rugs, leaving the reliability of the instrument in doubt.

In 1947 Battorf in "A Study Comparing Two Methods of Developing Art Appreciation with College Students" (3) developed a series of illustrations of rooms, lettering styles, and architecture with A, B, and C choices available. The A choice for the room item featured two simple "pleasing" room arrangements contrasted to (C) two very confused and intricate arrangements and (B) two rooms that might be considered "partly pleasing" and "partly confusing." Subjects with several years of college art training scored higher on the instrument than non-art trained persons.

Later instruments include a series of test items developed by Williams in 1967 (32) which were constructed by random spreading of colors on small cardboard blanks by a machine. Designs were matched to produce similar pairs. One of each pair was judged more aesthetically pleasing by
a group of forty-one experts. Artists and non-artists taking the test were asked to choose the most aesthetically pleasing design. One result of the study reported by Williams was the lack of significant variation in scores obtained from both artists and non-artists. Williams had predicted this result on the basis that ability in studio work did not necessarily determine the level of aesthetic sensitivity. Kieselbach (21) in 1955 developed a rather elaborate three-dimensional abstract design with moveable sections which was very similar to the Graves Design Test. The instrument used the variations in the design selected by subjects which differed from the configuration determined most pleasing by the panel of judges as the source of the score.

A marked departure from the research presented in the preceding paragraphs which used the naive to sophisticated judgment approach is the study by Brent Wilson entitled "The Relationship Between Years of Art Training and the Use of Aesthetic Judgmental Criteria Among High School Students" published in 1972 (34). Taking a slightly different tack is the 1975 study by Morris on "An Alternative Methodology for Researching Art Attitudes and Values" (23). The significant change in the direction of these studies is examining statements made by the subjects regarding art objects rather than simply determining the degree of agreement between their choices and a panel of judges.
The Morris study (23) is primarily a lengthy interview survey which seeks to describe responses to selected art objects by Appalachian mountain youths. Wilson, on the other hand, not only requested judgment regarding the aesthetic merit of art objects, but also ranked those statements in terms of frequency and level of aesthetic judgmental criteria.

Wilson used the aesthetic judgment categories he constructed to assess through personal interview the number and kind of evaluation statements made by non-art and art students with the hypothesis that persons with one or more years of art training would consistently make higher level statements in greater quantity than those without art training. Using Chagall's "I and the Village," each student was asked to rate the painting as "good, fair, or poor" and to give his reasons for the judgment. The transcript of each interview was analyzed to determine how many sentences contained evaluation criteria. Content analysis of each sentence was conducted to ascertain the level of criteria being employed. Scores for "each subject's use of each category were computed by determining the total number of sentences in the responses and calculating the percentage of the total responses taken by each category" (34, p. 30). The results indicated that Wilson's hypothesis was correct and further, that the categories (organistic, contextualistic, hedonistic, communication, originality, correspondence to reality, technical
achievement, moral worth, narrow a priori, authority, and personal preference) could be employed by expert scorers with a very high degree of inter-scorer reliability (34).

Interscorer reliability was determined by "a procedure that indicated the number of times there was total agreement between the two scorings" (34, p. 37). Wilson indicated some scorer reliability problems in the organistic category (formal elements) and in the broom closet (miscellaneous and unclassifiable responses). However, all other categories reported reliability scores ranging from seventy-one to ninety-nine percent. In his study Wilson made a distinction between weak and strong statements in each category. A weak statement would contain only vague general references while strong statements would be more specific and descriptive. The difficulty that the scorer encountered in determining those weak and strong differences was cited as a possible reason for the reliability problems with the organistic category. Wilson recommended abandonment of the weak and strong designations (34).

Problems and Possibilities of Aesthetic Judgment Tests

All of the instruments which belong to the comparison of naive to sophisticated judgment canon of research present difficulties. Most obvious is the limitation of the instrument to formal design qualities such as in Graves (19),
Thorndike (20), Gordon (18), Welsh (33), Battorf (3), and to some extent Meier (22). Broudy lists six aspects of a work of art in any medium that can be examined and evaluated:

(a) the sensuous materials; (b) the manipulation (technique); (c) the formal design; (d) the expressiveness or significance; (e) the general function of being interesting to perception and (f) the extra-aesthetic functions that it is designed or happened to serve (9, p. 95).

Basing an evaluation of aesthetic sensitivity on formal design elements alone severely limits the degree of validity that can be construed for such an instrument when it purports to measure the behavior of a subject toward an art object.

Those instruments utilizing reproduction of art works create different problems by assigning a standard of good and bad art that is considered to be constant even though the literature of aesthetics indicates many shifts in values (26). The Battorf instrument (3) is characteristic. The A choice of a pleasing simple room may well have sufficed in 1947, but the eclectic decorating style of the 1970's would probably find the A choice sterile and the least preferred C choice more aesthetically pleasing. The Child Test of Esthetic Sensitivity (11) and the Meier Art Judgment Test (22) are also examples of this difficulty. Shifting and rearranging masterpieces is one of the standard techniques in
Pop Art, thirty years after the development of the Meier Art Judgment Test.

The Beittel Art Acceptance Scale and the Christensen-Karwoski Test for Art Appreciation which include reasons for selections both share the possibility of influencing judgments by drawing attention to characteristics that might not otherwise be observed by the subject as well as eliminating any but predetermined responses (24, 33). As Broudy again states, "If we could predict the response, we would not be dealing with art or aesthetic experience, we would be doing empirical psychology" (8, p. 16). Osborne indicates that once attention has been called to aspects of an art work, the perception of the object changes, "as the intensity and manner of attention changes so the object of each man's perception is transformed . . . " (24, p. 14).

Finally, and most significantly, the scores produced by the instruments described above yield little or no information regarding the behaviors resulting in those scores. A result simply indicates the degree of agreement between the subject and the panel of experts. As indicated in Broudy's description of elements of art works, this agreement could result from any or a combination of the six aspects of art objects. Non-art trained students would probably score lowest and for a greater variety of reasons on these naive to sophisticated judgment comparisons.
Despite these deficiencies in the standardized and experimental instruments, they are still in frequent use in research projects. The Meier Art Judgment Test and the Graves Design Test were utilized in a 1960 study by Brandon concerning the "Relative Affectiveness of Four Different Approaches in Developing Art Appreciation" (7). Anderson in 1974 used the Child Test of Esthetic Sensitivity, the Graves Design Test and the Eisner Art Attitude Inventory in a "Comparison of Three Methodologies on Affecting Aesthetic Sensitivity" (2). The Eisner Art Attitude Inventory measures attitudes about involvement in art activities and is not a measure of aesthetic judgment (20).

Neither study reported significant differences among the various methods employed to develop art appreciation. It is conceivable that this is a legitimate result of the research. It is also possible, as Wilson notes in his critique of the Brandon investigation, "no necessary connection existed between what was tested in the Meier and Graves Tests and what was taught in the art appreciation courses" (33, p. 26). This possibility is underlined by Davis' appraisal of the standardized art instruments developed prior to the 1940's, a category which includes both the Meier and Graves instruments:

These early tests were hindered by the fact that there were no working definitions of the factors which were being measured. . . . The result was that many of the tests measured things which they
were not constructed to measure and did not measure factors which they were supposed to measure (13, p. 14).

The objectives of both the Brandon and Anderson studies (7, 2), to discover the most effective of several teaching strategies in affecting attitudes toward art, is certainly within the context of the goals of art appreciation as reported by Diffily (15) in his review of writings of art educators, art critics, aestheticians, and social critics in "Art Appreciation and Attitude Change: An Interdisciplinary Approach to the Development of Aesthetic Response in College Students." Diffily's study demonstrated that the factual and informational approach to teaching art appreciation will often reinforce negative attitudes toward art if no attempt is made to modify those attitudes through techniques of attitude change (15).

It is evident that the existing instruments for measuring aesthetic judgment are flawed by limitations to design elements, by pre-supposing standards of good and bad art, by pre-determining responses, and finally, by influencing perception of art works. It is evident that the scores produced by naive to sophisticated instruments yield insufficient information on the basis of an aesthetic judgment made by a subject. It is also evident that in order to affect basic aesthetic decisions and behavior which is crucial in developing
appreciation, the identification of the premise employed in an aesthetic judgment is exceedingly important.

In contrast to this lack of useful information from the standardized and experimental instruments described above, the Wilson study offers significant possibilities (34). Wilson made use of content analysis of student statements to determine the aesthetic criteria employed in evaluating an art work. Though his interest was primarily in the number and kind of judgmental statements made by the students, the process could be used to determine the predominant aesthetic criterion used by a viewer.

The ideal application of the Wilson methodology would be to conduct similar interviews with students on their appraisals of art objects and analyze the content of their responses in terms of the judgment categories. Unfortunately, the large number of students and the extreme limitation of time (16) precludes this possibility. Modification of both the data gathering process and the analysis of the responses will be necessary if the Wilson categories are to be useful in art appreciation courses.

Requirements for Aesthetic Judgment Instruments

The preceding critique of the available aesthetic judgment instruments provides both possibilities and problems
in the design of a new instrument. Other requirements for new instruments are available from the literature on test construction.

Among the requirements specifically related to aesthetic judgment instruments is the need cited by Child (11, p. 49) to replicate an actual judging situation by first requiring the subject to select the better of two possibilities and second by use of reproductions of art work rather than abstract designs which are not art objects. Child's discussion of an earlier aesthetic judgment instrument includes the following: "It asks for aesthetic judgment about actual works of art. Such a test has better initial claim to measuring aesthetic sensitivity than tests whose stimuli are not works of art and whose instructions do not direct attention to aesthetic values" (11, p. 49). The chief criticism of the Graves Design Judgment Test was its reliance on arrangements of formal design elements rather than art objects, excluding many other aspects of an art work (33, 35).

Another need indicated in criticism of the Beittel and Christensen-Karwoski instruments is avoidance of pre-recorded responses that eliminate the individual remarks of the subject and further draw attention to aspects of the art work that may or may not have been part of the subject's behavior toward the art object (34, 25).
In addition to these considerations which are directly related to the measurement of aesthetic judgment making, Karwoski and Christensen note the general necessity of using language which is readily understandable to the test subjects, a concern emphasized by Buros in constructing test manuals (12, 10). Buros also states that test instructions and administrative procedures need to be as uncomplicated and lucid as possible for the greatest reliability in using the instrument (10).

Finally, as indicated in the critique of the available instruments, a new aesthetic judgment test must serve its intended purpose to measure aesthetic judgment and must do so consistently. These two characteristics are generally referred to as validity and reliability, respectively (1, 24).

Validity, according to Ahmann and Glock, requires consideration of four distinct types: content, concurrent, predictive and construct (1, pp. 291-99). The critical factor in content validity is the "degree to which the instrument can be used as a basis for evaluating the present performance of the subject in a given situation" (1, p. 318). Content validity is generally determined by examination of the instrument in regards to the appropriateness of the method used to produce the measurement. Child's comments regarding replication of the aesthetic judgment situation
are particularly applicable in this consideration (11). Concurrent validity is primarily concerned with indirect measurement of skills or information. Predictive validity is demonstrated by the ability of the instrument to predict accurately future behavior in the area under scrutiny. Construct validity refers to the theoretical basis for the behavior that is being measured. If construct validity exists in the instrument, the trends in the behavior being tested should be predictable from the related literature (1, 24). In the area of aesthetic judgment making, concurrent validity might be evident in the Williams study that used selections of colored cardboard swatches to indicate aesthetic sensitivity in reacting to art works (32). Predictive and construct validity for an aesthetic judgment instrument would be concerned with determining the level of appreciation and its effect on selection of preferred art works. A subject with low aesthetic sensitivity, on the basis of the literature, would be expected to choose realistic and uncomplicated art works (8, 11, 25, 31).

The ability of an instrument to measure a given situation consistently, which is termed reliability, is generally investigated through determination of equivalence and stability (1). Equivalence requires similar instruments to produce similar results. Stability is based on the ability of an instrument to yield nearly identical results when given
to the same subjects twice with a brief time lapse between administrations. Equivalence is most often determined by correlating the results of two versions of an instrument or the results of the experimental instrument with a standardized instrument. The results of two administrations of the same or equivalent versions of the instrument are used to determine stability (1, 24).

Summary

The characteristics and relationship of appreciation and aesthetic judgment making have been considered as well as the problems and possibilities of the available methodology for measuring aesthetic behavior. Requirements for development of an aesthetic judgment instrument have been discussed in terms of the particular concerns in aesthetic behavior and the general need for valid and reliable instruments. Based on the lack of adequate instruments, a modification of the Wilson study with its aesthetic judgment categories has been proposed as a possible source for a new aesthetic judgment instrument.
CHAPTER BIBLIOGRAPHY


CHAPTER III

DESIGN OF THE STUDY

Introduction

Given the valuable information for the art educator available through an interview survey to determine aesthetic judgment premises as demonstrated by Wilson (13) and the high degree of reliability established for the categories of aesthetic judgments in the same study coupled with the deficiencies in information and validity of the existing instruments in aesthetic judgment, it seemed worthwhile to construct an instrument based on the aesthetic judgment categories which also reduced the amount of time necessary to obtain the required information. Eliciting a brief written response stating reasons for a choice had distinct possibilities because of the readiness with which people express opinions on the merit of art works (10, 12). The possibility of the subject analyzing his own responses and selecting the appropriate category from the Wilson group was also viable from the reported results of the Beittel and Christensen-Karwoski instruments which contained reasons for choices (3, 5). The combination of open-ended response and self-analysis of the response could avoid the problem
of pre-determined responses which eroded the usefulness of both the Beittel Art Acceptance Scale and the Christensen-Karwoski Test for Aesthetic Appreciation. Among the significant possibilities for such an instrument would be (1) rapid identification of entering aesthetic judgment behavior; (2) subsequent adoption of activities and teaching strategies to modify specific behaviors; (3) precise evaluation of change in those behaviors, and (4) a more sensitive and specific method for measuring the effectiveness of one teaching strategy against another.

Statement of Purpose

The purpose of this study was to provide an instrument which elicits criteria for aesthetic judgments and assigns those criteria to the Wilson categories which are ranked by degrees of adequacy.

Statement of the Problem

The problem of this study was to develop an instrument which differentiates levels of aesthetic judgment making in the categories derived by Wilson and which could be scored rapidly with a high degree of reliability.

Development of the Instrument

History

Publication of the Wilson study (13), "The Relationship between Years of Training and the use of Aesthetic Judgment
Criteria among High School Students," and a developing concern over the diversity of the students enrolling in the art appreciation classes occurred, fortuitously, in the same fall semester of 1972. The categories were immediately recognized as a means of bringing orderly classification to the myriad statements made by students regarding art works. They also provided a possible solution and to extremely puzzling instructional problem. All too frequently, students, after considerable instruction in the process, were satisfactorily completing critiques of art works using the Feldman model of description, analysis and interpretation. When reaching the evaluation phase, however, the data developed from the analysis were totally ignored and statements such as "It is a good work of art because the subject matter has a strong message about human freedom" were written.

Analysis of this difficulty, in light of the Wilson categories, indicated that the premise on which an aesthetic decision was made was usually not affected by learning the elements and principles of art and their application unless the premise was first examined for adequacy or inadequacy. This need for information prompted the first version of this instrument, a simple request to state in writing whether or not a work of art was good and to give the reason for the evaluation.

The written statement has been a part of the instrument from its inception. The possibility of conducting interviews
with each student and analyzing the statements was rejected when the amount of time involved was calculated. Twenty minute interviews would have required an excessive amount of instructional time to gather data. In order to confront and modify an inadequate judgment premise, the information was needed at the very outset of the semester. Other test possibilities such as the Meier or Beittel instruments had already been rejected because of their tendency to contaminate and change initial perceptions of an art work as well as not revealing sufficient information (8, 13, 14).

The possibility of using a self-scoring component occurred as a result of a group class assignment to select winners in an art show. An additional problem in the assignment was to select from a list of possible reasons the one that most accurately indicated the group's basis for selection of one art work over another one. Success with this venture indicated that students could make some judgment about the meaning and intent of their reasons for selections.

Formal development of the evaluative instrument began after the review of the literature on aesthetic judgment. An added impetus for development was the need to devise a method of evaluation which could be used in satellite testing centers with large numbers of students on each of five community college campuses and which could be scored by computers. It was apparent that an instrument which met these
conditions also be applicable to the large sections of non-major students that Feldman (6) had described.

The review of the literature indicated that the following guidelines would need to be met:

1. The aesthetic judgment making situation would need to be replicated through directions to make a judgment regarding the merit of art works.

2. The stimuli for the aesthetic decision would have to be highly accurate color reproductions rather than abstract designs or some other non-art illustration.

3. Directions would have to be in readily interpretable language.

4. Administration procedures would need to be as simple as possible.

5. Avoiding pre-determined responses would necessitate an opportunity for an open-ended response.

6. Avoiding contamination of that response would require that any self-scoring device would be withheld until the written response was completed.

As a result of these findings, care was taken to ensure that students were asked directly to make a judgment about color slides of paintings or sculpture. The Wilson categories were rewritten to remove the formal language of aesthetics. Examples of student statements which had been placed in the category during the informal field testing
were substituted for terms such as organic, unity, and intrinsic formal value. Thus the mimetic category, which was defined as "correspondence to reality," became "it looks like the real thing."

The major difficulty inherent in an instrument that is self-reporting in nature is that the data will be as reliable as the perceptions of the individual reporting the data (9). Precautions were taken in the arrangement of the categories in the self-scoring component to avoid any indication of preferred premises. However, the lack of reluctance by most people to state opinions about art works as chronicled by Osborne (10) and in the UNESCO study (12) was considered an advantage in this instance. The "I'm not an expert" syndrome seemed to promote a decided honesty in the written statements.

Description of the Instrument

Meeting the guidelines discussed above resulted in a two-part instrument (Appendix A). Part one consisted of viewing a pair of slides of art works with the following instructions:

The two art works showing on the screen are both by well known and respected artists. Select the work you consider to be the best. Mark your answer below:

The best work is the one on the left. □

The best work is the one on the right □

In the space below, write your reasons for your choice. Please use complete sentences.
Subjects were required to complete the statement of reasons before receiving part two of the instrument. Titles for the art works were not given.

Part two of the instrument, the self-scoring component, listed the Wilson categories in random order. Each category was represented by a series of typical statements, one of which said, "Good art is _____" and contained the primary criterion for the category. The instructions for this section were as follows:

Listed below are a number of reasons that people give for deciding one art work is better than another. Look over the statement you have just written. Then, from the groups of comments printed here, select the one group which seems closest to the reason you have given. You probably won't find an exact match in wording, but you should find statements with a close match in meaning. Choose only one group.

For example, if you have decided that the painting on the left is better because it looks more like the real thing, you would choose the group of statements labeled B.

When you have decided which group of statements is right for your reasons, put the letter for that group in the box in the right hand corner of the page on which your statement is written.

**Administration Procedures**

Testing time averaged approximately fifteen minutes. Two slides of recognized art works, chosen on the basis of the guidelines which follow, were projected on the screen
at the front of a darkened classroom. As each subject completed the written statement (part one), the self-scoring component (part two) was distributed individually. When all students had completed part one, the slides were turned off as a check against ignoring the written statement as the basis for a category choice.

Guidelines for Selection of Art Works

No significance was attached to which slide was selected in part one of the instrument. The two art works were used to provide an appropriate focus for making an aesthetic judgment and, further, to limit whatever definitions of art existed in the subject's considerations. Because the intent of the instrument was to elicit the basis for the choice rather than comparison of naive to sophisticated judgments, both art works were recognized masterpieces. In addition to being accurate color reproductions, the characteristics and content of the slide pair had to provide opportunity for utilization of any of the Wilson categories in making aesthetic judgments. This requirement made inclusion of recognizable subject matter necessary in at least one of the slides to satisfy the moral worth, mimetic, and communication categories. A reasonable element of novelty in arrangement or content of the subject matter to accommodate the originality category was included. Familiarity of the image, the artist's style, or the subject matter sufficed for the authority
category. Some attention to detail opposed to a more painterly or generalized technique in the slide pair met the requirement for technical achievement. Finally, selection of one work which emphasized formal considerations in contrast to one which was primarily expressive provided for, respectively, the organistic category and the contextualistic and hedonistic categories.

In preparation for the studies of reliability and validity, three pairs of slides were selected to produce Version A, Version B, and Version C of the instrument. The primary difference among the pairs was the degree of abstractness present in one work within each group of two slides.

**Description of the Slide Pairs**

Version A, Picasso's "Nudo Seduto" and Matisse's "Figure in an Ornamental Background" consisted of two female nudes. The Picasso is basically a colored line drawing of a solemn, older, and rather frail woman seated on the edge of a bed. This image uses somber colors in shades of pale red and blue. The overall effect is one of quiet resignation. The Matisse places a recognizable, if somewhat geometric, seated nude in the midst of a very decorative, almost flamboyant abstract room in which the patterns of wall paper and carpeting merge.

Version B, Van Gogh's "Cornfield with Cypresses" and Chagall's "I and the Village" are essentially two landscapes.
The Van Gogh, done in bold yellow and black, is fairly straightforward with vigorous and characteristic brush-strokes forming the pattern of the corn in the lower half of the painting. The upper half of the image is devoted to an intensely blue sky. A flamelike cypress tree dominates the right hand side of the painting. Chagall's "I and the Village" is precisely rendered and realistic. However, the many separate images in the composition are arranged in an abstract pattern with the head of the artist and the head of a cow opposing each other on each side of the vertical canvas. These two shapes serve as a framework for several small representations of the artist's childhood, all carefully executed vignettes.

Version C, Rembrandt's "Syndics of the Cloth Guild" and Giacometti's "People in the Piazza" are both images of groups. The Rembrandt is a very solemn portrayal of the guild members in black and brown with careful attention to the portraits of individual group members. The gentlemen are seated in a half-circle around a rug-covered table with a large book laid open on it. In contrast, the Giacometti is a metal sculpture with a flat rectangular base from which six very attenuated and generalized slender stick figures emerge at varying heights. Each elongated figure has a similar over-large but still narrow head.
Considerations in Determining Validity and Reliability

Ahmann and Glock list four types of validity and two kinds of reliability to be considered in the assessment of the qualifications of an instrument. The four types of validity are content, concurrent, predictive and construct. Reliability is considered in terms of equivalence and stability (1, 9).

The instrument described in the preceding paragraphs was designed to reveal the criteria used in making an aesthetic judgment and further, to categorize those criteria by degrees of adequacy through placement in the Wilson list of aesthetic premises. The instrument presented a pair of art objects and asked the subject to select the better of the two works and to list the reasons for that decision. After listing the reasons, the subject was then asked to select one set of statements out of a list of categories of statements that most closely matched the meanings of the reasons he gave.

Replication of the circumstances in which aesthetic judgments are made was crucial in consideration of content validity for this instrument. The use of a direct request to make a judgment about two works of art met the requirements established by Child in considering validity (4) for aesthetic judgment instruments. Further, validity for the
judgment categories was ascertained from the number and kind of statements that had to be classified in the miscellaneous or broom closet category. These statements would be significant if they indicated that the categories developed by Wilson failed to accommodate the majority of the aesthetic statements made by the subjects. If, however, the majority of statements elicited by the instrument could be categorized by the Wilson premises, then the purpose of the instrument in terms of both eliciting aesthetic criteria and ranking that criteria would be met.

The preceding discussion of content validity negates consideration of concurrent validity for this instrument. The instrument was not concerned with indirectly measuring aesthetic judgmental criteria, but rather assessed those criteria directly through instructions to make an aesthetic judgment regarding two art works.

Based on the evidence reported in the UNESCO study regarding reactions to art works (12), identifying the aesthetic premise employed by a subject makes inference of a degree of predictive validity possible for this instrument. This is particularly true for those categories designated as inadequate or somewhat adequate. Use of the criteria of correspondence to reality and readily grasped meaning resulted in the UNESCO respondents rejecting forty of the major painters of the twentieth century including Picasso,
Klee, Mondrian, Pollock, and Miro (12). It is reasonable to expect that subjects whose premise for aesthetic judgments preclude consideration of the art work in its totality would assign less aesthetic merit to works which did not meet their pre-determined prerequisites. Demonstration of this implied ability to predict behavior toward art works was beyond the scope of this study.

The variations in levels of adequate and inadequate judgment criteria appeared to be the result of a construct that has been defined as aesthetic sensitivity: "It refers to the extent to which a person gives evidence of responding to relevant stimuli . . . " (4, p. 49). These relevant stimuli as described by Ziff may include, "what kind of work it is, who painted it, what kind of organization it has, what kind of subject matter (if any), what kind of pigmentation, . . . " (15, p. 234). Considering the evidence present in the work itself such as the items just listed was the determining factor in designating aesthetic criteria as adequate. Inadequate criteria ignored some or all of the possible aspects of the art work (11).

Considerable evidence resulting from studies designed to measure aesthetic sensitivity indicates that the construct of aesthetic sensitivity is affected by training. For example, Wilson, Battorf, and others (13, 2) predicted that persons with college art training would score higher on aesthetic
sensitivity than persons without art training. As reported by these studies, the results have corresponded with those predictions. Meier, in his test for art judgments, used the higher scores of art trained students as one of his measures of validity (8). As required in the determination of the validity of a construct, it is possible to predict trends in the behavior being measured from the theory available in the literature.

Construct validity is usually confirmed through comparing the results of equivalent instruments (1). The lack of any other instrument concerned with determining the premise of an aesthetic judgment effectively eliminated this comparison possibility.

The ability of an instrument to measure a given situation consistently is generally investigated in terms of stability and equivalence (1, 9). Equivalence is usually established by administering two similar instruments and correlating the results. At present, no instrument equivalent to the one developed for this study exists. Further, the possibility of using one of the naive to sophisticated judgment comparisons as an equivalent instrument was eliminated in the critique of the instruments which indicated that no source for a low or high score could be determined. The impetus for the development of this aesthetic judgment instrument was a result of this lack of information regarding
the causes of scores obtained on the comparison tests. No research evidence exists for equating a high score on a comparison instrument with one of the most adequate Wilson categories. The only alternative was to vary the composition of the slide pairs to produce somewhat different stimuli for judgment making.

Stability for an aesthetic judgment instrument also presented a problem. As Osborne (10), indicated, drawing attention to any aspect of an art work would tend to change the viewers perception of the art work. The impact of this tendency on the commonly accepted procedure for determining stability which is to administer the same instrument to the same group of subjects in a test-retest pattern, allowing a two to three week lapse of time between administrations (1, 9), was one of the factors to be considered in the reliability study. As a safeguard against this tendency the instructions devised for the self-scoring component clearly state to use the written statement as the source for selecting a category, and the administrative procedures require withholding access to the categories until the first statement is completed.

Research Design for Reliability and Validity Studies for the Instrument

The research design developed for this study of reliability and validity for the instrument was a test-retest
pattern with comparison of (a) results of scoring and re-scoring by an expert scorer on the first administration; (b) results of expert scoring and self-scoring on the first and second administration; (c) results of self-scoring on the first and second administration, and (d) results of expert scoring on the first and second administration.

Three versions of the instrument were used. The versions differed only in the composition of the slide pairs. Results for all four comparisons listed above for the three forms of the instrument were compared to determine their equivalence. In addition, the responses ranked in the "broom closet" were examined to tabulate both the number and kind of statements included to determine the degree of content validity.

Hypotheses

The research hypotheses formulated for this study were as follows:

1. There will be a high percentage of agreement between ranking of individual responses by the expert scorer and rescoring of the same responses by the scorer after a three-week interval.

2. There will be a high percentage of agreement between the ranking of individual responses produced by the self-scoring component and ranking by the expert scorer in the first and second administrations of the instrument.
3. There will be a high percentage of agreement on rankings of individual responses produced by the self-scoring component on the first and second administration of the instrument.

4. There will be a high percentage of agreement between the ranking of individual responses by the expert scorer on the first and second administration of the instrument.

5. The percentage of agreement for the three versions of the instrument will differ by fewer than ten percentage points for all four comparisons.

6. Less than ten percent of the total responses scored by the expert scorer will contain no aesthetic criteria and will be classified in the "broom closet" category.

Limitations of the Study

The following limitations were imposed on this study:

1. The study was limited to construction of the instrument and tests for its validity and reliability.

2. The utilization of the instrument was limited to post-secondary school populations.

Assumptions

For this study the following assumptions were made:

1. It was assumed that the categories developed by Wilson are sufficiently comprehensive to include the possible range of aesthetic judgment criteria.
2. It was assumed that, as stated by Pepper (11), the most adequate judgmental criteria are those that are based on the work of art.

3. It was assumed that the less adequate judgmental criteria are pre-determined standards that originate outside the art object.

Definition of Terms

The aesthetic judgment categories utilized in this study were those developed by Wilson (13) from an extensive review of the literature in aesthetics. They were defined as follows:

1. Organistic: the criterion of unity, coherence, cohesiveness and complexity.

2. Contextualistic: the criterion of intensity and vividness.

3. Hedonism: the criterion of affective pleasure or pain.


7. Mimesis: the criterion of correspondence to reality.

8. Technical achievement: the criterion of skill.

9. Narrow a priori: the criterion of pre-determined features.
10. Authority: the criterion of expert opinion.
11. Personal preference: the criterion of what is liked.

Description of the Sample
The subjects included in this study were students who enrolled in two sections of day-time freshman general math classes on each of three separate campuses of a major metropolitan community college district in the spring of 1977. Math classes were chosen to prevent contamination of the sample by possible encounter with art judgment considerations as a part of the curriculum. A total of one hundred fifty-five students were present of the first administration of the instrument. Ninety-two students were present for the second administration three weeks later. Students whose responses to a brief questionnaire indicated current enrollment in art appreciation or humanities survey courses were eliminated from the population of the study to produce these totals.

Procedures for Collection of the Data
Prior to the first administration of the instrument, students completed a brief questionnaire that required the following information to be used as a source document for recording and processing data: social security number and
current enrollment in an art appreciation or humanities survey course. A three-week lapse of time separated each administration of the test versions. All tests were given in similar air-conditioned college classrooms seating approximately thirty-five persons. Test versions for each group were assigned on a random basis and were given during the regularly scheduled morning class period.

Version A (Picasso-Matisse) was given to forty students during the first testing period and twenty-six students in the second administration. Sixty-three students were given Version B (Van Gogh-Chagall) in the first administration and twenty-seven were present for the second test period. Version C (Rembrandt-Giacometti) was administered to fifty-two students during the first test period and thirty in the second administration.

Interscorer reliability was established by the expert scorer, a college art appreciation instructor with considerable training in aesthetic judgment-making, scoring and rescoring the first set of written responses on part one of the instrument with a three-week time lapse between scorings without knowledge of the results of the self-scoring component.

Procedures for Treatment of the Data

Six comparisons of results for each version of the test were conducted to determine the reliability and validity of the instrument. The six comparisons were as follows:
1. Expert scoring and rescoring of the results on the first administration for all test versions (Interscorer reliability).

2. Expert score and self score on the first administration for all test versions (Self-scoring component reliability).

3. Self-score on the first and second administrations for all test versions (Self-scoring component stability).

4. Expert score and self-score on the second administration for all test versions (Self-scoring component reliability).

5. Expert score on first and second administrations for all three test versions (Instrument stability).

6. Tabulation of the number and kind of responses placed in the "broom closet" category by the expert scorer (Content validity).

Borrowing methodology from the Wilson study (13), the results of each of the five comparisons for each of the three test versions were expressed as the percentage of exact matches of category choices. Scores on the test consist of letters signifying the category in which a response has been placed. The categories have no numerical dimensions. In order to determine the percentage of agreement, the number of direct matches were compared to the total number of responses on the test. For example, exact agreement between
the expert scoring and the results of the self-scoring on
the first administration of Version A in twenty-nine of
the sixty-three responses would result in a percentage of
agreement of forty-eight percent. The instrument, there-
fore, did not result in direct matches between categories
in thirty-four instances.

The Wilson study reported interscorer reliability as
acceptable in percentages of agreement ranging from seventy-
one to ninety-nine percent. For the purposes of this study
a percentage of agreement of eighty percent seemed reasonable
for all comparisons.
CHAPTER BIBLIOGRAPHY


CHAPTER IV
FINDINGS AND INTERPRETATION OF THE DATA

Introduction
A test-retest design was employed for three versions of the aesthetic judgment making instrument with a self-scoring component. Each of three groups of freshman math students were given one version of the test twice with a three-week lapse of time between administrations. One hundred fifty-five students were present for the first administrations. Ninety-five students participated in the second testing period.

Comparisons for interscore and interscorer reliability were conducted for each of the three test versions. Overall results among the three versions were compared to determine the degree of equivalence of the instruments. The number and kind of responses placed in the "broom closet" category by the expert scorer were tabulated for content validity evidence.

Research data were reported as percentages of exact matches of categories out of the possible number for each group of test results. Based on the percentages of agreement reported in the Wilson study (3) which ranged from
seventy-one to ninety-nine percent, a percentage of agreement of eighty percent was deemed acceptable.

Hypothesis One

Hypothesis One was stated as follows: There will be a high percentage of agreement between ranking of individual responses by the expert scorer and rescoring of the same responses by the expert scorer after a three-week interval. Included in Table I is the report of the results for the three test versions

TABLE I

PERCENTAGES OF AGREEMENT FOR EXPERT SCORING AND RESCORING OF TEST VERSIONS A, B, AND C

<table>
<thead>
<tr>
<th>Test Version</th>
<th>N</th>
<th>Number of Matches</th>
<th>Percentage of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40</td>
<td>39</td>
<td>.98</td>
</tr>
<tr>
<td>B</td>
<td>63</td>
<td>58</td>
<td>.92</td>
</tr>
<tr>
<td>C</td>
<td>52</td>
<td>47</td>
<td>.90</td>
</tr>
</tbody>
</table>

The hypothesis was retained on the basis of the extremely high percentage of exact matches for scoring and rescoring by the expert scorer. The high interscorer reliability matches the result of the Wilson study that formed the base for this instrument, adding credibility to the categories for content analysis. The Wilson study reported reliability percentages
of agreement ranging from seventy-one to ninety-nine percent. For this instrument, the interscorer reliability range for the three test versions is ninety to ninety-eight percent.

Hypothesis Two

Hypothesis Two was stated as follows: There will be a high percentage of agreement between the ranking of individual responses produced by the self-scoring component and ranking by the expert scorer on the first and second administrations of the instrument. Table II contains the result of the self-score and expert score for all test versions and for both administrations.

<table>
<thead>
<tr>
<th>Test Version</th>
<th>N</th>
<th>Number of Matches</th>
<th>Percentage of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>40</td>
<td>14</td>
<td>.35</td>
</tr>
<tr>
<td>B</td>
<td>63</td>
<td>29</td>
<td>.46</td>
</tr>
<tr>
<td>C</td>
<td>52</td>
<td>31</td>
<td>.60</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Second Administration</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
</tr>
<tr>
<td>B</td>
</tr>
<tr>
<td>C</td>
</tr>
</tbody>
</table>
The marked increase in agreement between expert scoring and self-scoring in the second administration for Version A and B and the slight decrease for Version C caused rejection of Hypothesis Two in line with the contingency discussed under considerations for stability. As described by Osborne (2), the contingency was that directed encounters with art works tend to change perception of the work. In this instance, the instrument is apparently serving a teaching function in terms of vocabulary. The expert scorer noted that particularly for Version A, wording used by students in the second test statement showed considerable commonalities with the category statements used in the self-scoring component. Test memory (1) was probably a factor in this occurrence as well as the possibility of the instrument supplying new terminology and perceptions for the students. The effect of other variables such as the decrease in the number of students present for the second test period could not be determined.

The possibility of the convergence of wording continuing in subsequent administrations of the instrument will need further investigation. However, the generally low percentages of agreement between expert scoring and self-scoring on the first administration point to two possible problems: (1) the student ignored the written statement and selected a category that corresponded to his intended if unwritten premise, or (2) content analysis of the written statement may
require a skill and objectivity not possessed by these students. Problem one, in particular, will need to be investigated in another study. It is quite possible that the selected statement is a more accurate self-report than a written statement.

Hypothesis Three

Hypothesis Three was stated as follows: There will be a high percentage of agreement on rankings produced by the self-scoring component on the first and second administration of the instrument. Table III contains the data for this comparison.

TABLE III

PERCENTAGES OF AGREEMENT BETWEEN RESULTS OF THE SELF-SCORING COMPONENT FOR TEST VERSIONS A, B, AND C IN THE FIRST AND SECOND ADMINISTRATIONS

<table>
<thead>
<tr>
<th>Test Version</th>
<th>N</th>
<th>Number of Matches</th>
<th>Percentage of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35</td>
<td>20</td>
<td>.57</td>
</tr>
<tr>
<td>B</td>
<td>27</td>
<td>13</td>
<td>.48</td>
</tr>
<tr>
<td>C</td>
<td>30</td>
<td>3</td>
<td>.10</td>
</tr>
</tbody>
</table>

On the basis of approximately half of the scores matching for Version A and B and only ten percent of test version C results agreeing, the hypothesis was rejected.
These results indicate that the self-scoring component is virtually unreliable. Part of the cause is most probably the tendency of perception of the viewer to change when attention is drawn to aspects of an art work. Version C of the test used the Rembrandt-Giacometti pair of slides, both of which were group portraits. Of all three test versions, this pair of slides had the greatest degree of difference with the highly realistic Rembrandt on one end of the scale and the very abstract Giacometti on the other. It can be conjectured that the student who in the first administration selected the Rembrandt on the basis of realism may have been more attuned to the composition or the impact of the work on the second test after encountering those dimensions of responding to an art work while using the self-scoring list of categories.

Examination of two sets of responses selected from the group of students viewing the Rembrandt-Giacometti pair may provide some insight into this occurrence as well as indicate some of the difficulties encountered by the expert scorer and the subsequent results for the instrument's reliability.

Subject A selected the Rembrandt on the first administration and gave the following rationale:

I enjoy more romantic or traditional forms of art. The left picture has actual colors and distinct figures. I can relate to it more than the abstract sculpture. The facial expressions interest me and I find that there is more to enjoy in viewing the left painting.
The expert scorer classified this response in the hedonistic category as a result of the stress on feeling and enjoyment. Though some formal elements such as "actual color and distinct figures" are mentioned, they seem to be secondary in importance.

Subject A, however, classified this statement in the mimetic category which in the self-scoring component included: "It looks like a photograph. It is very realistic. It looks like the real thing. Good art should look like something." It is possible that the statement on "actual color and distinct figures" was meant to convey realism. It is possible, also, as indicated in Hypothesis Two, that the student found a more accurate statement of her concerns in the self-scoring section and chose a new category.

In the second test period Subject A again chose the Rembrandt and gave the following reasons:

The painting on the left has color and clarity. It is an actual representation of a scene. It has more detail and offers more for a viewer to gain without an in depth knowledge of art.

The expert scorer placed this response in the organistic category because of the emphasis on color, clarity, detail, and interest. The response seemed to indicate more than mere realism and appeared concerned with formal elements. There is no mention of enjoyment or feeling, a marked departure from the earlier response.
Subject A selected the craftsmanship category which was illustrated in the self-scoring component with the following statements: "It looks like the artist knew what he was doing. It has taken a lot of time. The painting is done well. Good art should show that a lot of care has been taken with it." Close examination of the statement by Subject A reveals no basis for this interpretation.

In contrast, Subject B, while still selecting the Rembrandt painting, categorized the following statement in the mimetic category: "I liked the one on the left because there is more to see. It is also more colorful and much more realistic." The expert scorer agreed with the category choice.

The second administration for this subject produced the following statement: "It's far more interesting than the other. I believe more time and effort went into it," which was placed in the craftsmanship category by both the subject and the expert scorer. Though accurately scored in both administrations for this subject, the instrument produced different results.

The questions not pursued in this study, as indicated by both hypotheses two and three, were the causes of these results. Discovering the impact of changes in perception as demonstrated by Subject A's selection of new categories
and Subject B's report of different reasons need further investigation.

Hypothesis Four

Hypothesis Four was stated as follows: There will be a high percentage of agreement between the rankings of individual responses by the expert scorer in the first and second administration of the instrument. Data relating to this comparison are reported in Table IV.

<table>
<thead>
<tr>
<th>Test Version</th>
<th>N</th>
<th>Number of Matches</th>
<th>Percentage of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>35</td>
<td>16</td>
<td>.46</td>
</tr>
<tr>
<td>B</td>
<td>27</td>
<td>17</td>
<td>.63</td>
</tr>
<tr>
<td>C</td>
<td>30</td>
<td>13</td>
<td>.43</td>
</tr>
</tbody>
</table>

The failure of the instrument to produce similar written statements already chronicled in the discussions of hypotheses two and three is also evident in the low percentages of agreement achieved by the expert scorer. On the basis of the data reported in Table IV, the hypothesis was rejected.
Hypothesis Five

Hypothesis Five was stated as follows: Percentages of agreement for the three versions of the instrument will differ by fewer than ten percentage points for all four comparisons. A compilation of all comparison data is presented in Table V.

TABLE V

AMOUNT OF DIFFERENCE BETWEEN PERCENTAGES OF AGREEMENT FOR THE THREE TEST VERSIONS IN FOUR SCORING COMPARISONS

<table>
<thead>
<tr>
<th>Type of Comparison</th>
<th>Test Version</th>
<th>Percentage of Agreement</th>
<th>Amount of Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert score to rescore</td>
<td>A</td>
<td>.98</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.92</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>.90</td>
<td></td>
</tr>
<tr>
<td>Self-score to expert score</td>
<td>A</td>
<td>.35</td>
<td>Test/Retest</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.46</td>
<td>Test/Retest</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>.60</td>
<td>Test/Retest</td>
</tr>
<tr>
<td>Expert score to expert score</td>
<td>A</td>
<td>.46</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>.43</td>
<td></td>
</tr>
<tr>
<td>Self-score to self-score</td>
<td>A</td>
<td>.57</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>.48</td>
<td></td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>.10</td>
<td></td>
</tr>
</tbody>
</table>

Hypothesis Five was designed to produce evidence of equivalence for the three versions of the instrument. The three versions of the test cannot be considered equivalent on the basis of the data reported in Table V with the exception of the scoring and rescoring of the first set of
responses by the expert scorer. On the basis of the data, Hypothesis Five was rejected.

Hypothesis Six

Hypothesis Six stated: Less than ten percent of the total responses scored by the expert scorer will contain no aesthetic criteria and will be placed in the "broom closet" category. Thirteen responses of the original one hundred fifty-five were classified as "broom closet" responses. Of these thirteen, four contained no aesthetic criteria, and the remaining nine employed an aesthetic criterion not included in the Wilson categories. This criterion was somewhat related to the technical achievement premise in its concern with the amount of effort involved in the art work. However, the effort mentioned in the new criterion was on the part of the viewer rather than the artist. The criterion was expressed in these terms: "The painting really challenged my thinking. I had to work to figure it out."

The nine "broom closet" responses which contained the new criterion represented eight percent of the total responses. The four responses containing no aesthetic criterion equalled three percent of the total. As a result, the Wilson categories could be used to classify ninety-two percent of the total one hundred fifty-five responses. Further, the instrument produced aesthetic criteria in the written statements
in ninety-seven percent of the total responses. On the basis of these data the hypothesis was accepted.

Findings

The major findings resulting from this study are summarized as follows:

1. Interscorer reliability for the expert scorer ranged from ninety to ninety-eight percent in scoring and rescoring the first set of responses on the three test versions.

2. The self-scoring component of the instrument produced inconsistent percentages of agreement as a result of inaccurate placement in the categories by the subjects, changes in the subjects reasons for aesthetic decisions between the first and second administrations of the instrument, and selection by the subjects of new categories unrelated to the written statement.

3. Expert scoring of responses from administration one and two for the three test versions was also unreliable as a result of the failure of the instrument to produce similar written statements in both administrations.

4. Versions A, B, and C of the instrument which differed only in the composition of the slide pairs did not produce similar results and cannot be considered equivalent instruments on a test-retest basis.
5. The instrument produced aesthetic criteria in ninety-seven percent of the one hundred fifty-five initial responses.

6. Ninety-two percent of the aesthetic criteria produced by the instrument were classified in the aesthetic judgment categories.

7. Eight percent of the total one hundred fifty-five responses contained a new criterion related to the viewer's effort in understanding an art work.
CHAPTER BIBLIOGRAPHY


CHAPTER V

SUMMARY, CONCLUSIONS, IMPLICATIONS, AND RECOMMENDATIONS

Summary

This study was concerned with the development and testing of an instrument to measure levels of aesthetic judgment making. The review of related literature demonstrated that the majority of instruments for aesthetic judgment developed between 1896 and 1976 employed a naive to sophisticated judgment comparison to determine levels of aesthetic sensitivity. The inadequacy of a score reporting only the degree of agreement between the subject's choice and the choice of a panel of experts without indicating the source of agreement was discussed. Two recent research studies by Wilson and Morris which stressed content analysis of aesthetic responses were discussed as an alternative means for determining aesthetic criteria.

Findings from the review of related research in measurement of aesthetic judgment resulted in a two-part instrument. Part one required the subject to select the better of two art works and to state the reasons for the choice. Part two, a self-scoring component, consisted of the Wilson categories presented as typical statements containing the primary
criterion for the category. The subject was instructed to select the statements that were closest in meaning to his written statement of reasons completed in part one of the instrument.

Following a presentation of validity and reliability considerations, a research design on a test-retest pattern was constructed. Three versions of the instrument, which differed only in the composition of the slide pairs, were given to three groups of math students enrolled in a large metropolitan community college district. Each group was given one version of the test twice with a three-week time lapse between administrations. One hundred fifty-five students were present for the first administration, and ninety-five students were present for the second testing period. Students currently enrolled in an art appreciation or humanities survey course were eliminated from the sample before results were tabulated.

Comparisons for interscore and interscorer reliability were conducted for all three test versions. Overall results produced by the instruments were compared to determine the degree of equivalence for the instruments. Finally, the number and kind of responses placed in the "broom closet" category by the expert scorer were reported for content validity evidence.
Research data were reported as percentages of agreement, the number of exact matches in categories compared to total responses, between results for each set of scores on each of the three versions of the instrument. Hypotheses were analyzed and the findings are summarized in the following paragraph.

The categories for aesthetic judgments developed by Wilson produced percentages of agreement for expert scoring and re-scoring of the first set of written statements ranging from ninety to ninety-five percent, a result corresponding to the interscore reliability reported in the Wilson study which served as the basis for this instrument. The self-scoring component produced very unreliable scores with (1) student inability to select appropriate categories; (2) students choosing categories unrelated to the written statement; and (3) students changing reasons for aesthetic decisions between the two administrations as the probable causes. The inability of the instrument to produce consistent written statements affected all comparison scores negatively. The three test versions could not be considered equivalent as a result of differences in percentages of agreement ranging from eight to forty-nine percent. The Wilson categories were found sufficiently comprehensive to classify ninety-two percent of the total responses. One new category related to the amount of work required by the viewer to comprehend the
paintings was discovered in the examination of the responses classified in the "broom closet category." Ninety-seven percent of the initial one hundred fifty-five responses to the instrument contained aesthetic criteria.

Conclusions

The major conclusions derived from the findings of this study are as follows:

1. The Wilson aesthetic judgment categories consistently produce extremely high interscore reliability results.

2. The instrument developed for this study possesses a great degree of content validity on the basis of success in eliciting aesthetic criteria and classifying the criteria.

3. A test-retest research design for stability for an aesthetic judgment instrument is probably not an appropriate technique as a result of the tendency of directed encounters with art works to change perceptions of the work by the viewer.

4. The self-scoring component devised for this instrument contributed to its unreliability and was not successful as a self-analysis instrument for classifying aesthetic premises.

5. The instrument proved to be unreliable in terms of both stability and equivalence.
Implications

The implications for this study as derived from the findings are as follows:

1. This initial attempt to develop an aesthetic judgment instrument utilizing the Wilson categories demonstrates that the content analysis approach to determining aesthetic premises is feasible.

2. The success of the instrument in eliciting aesthetic criteria indicates that further modifications in scoring could prove useful in developing methodology for determining aesthetic premises.

Recommendations for Further Research

This initial attempt to develop a more valid instrument for evaluation of aesthetic judgment has raised a number of research questions, suggestions, and possible applications both as a result of the reported data and the survey and critique of available aesthetic judgment instruments in the review of related literature. These recommendations can be categorized under the following headings: further testing procedures for the instrument, additional research questions and observations related to aesthetic judgment premises, possible applications of the instrument in art curricula and teaching strategies, and suggestions to other investigators in this field of research.
Further Testing Procedures

Further investigation of stability and equivalence for the instrument can be conducted in a number of ways.

1. Because of the unverifiable but implied effect that exposure to reasons for judgments had on the performance of the self-scoring component in scoring responses, the instrument should be administered on test-retest basis without use of the self-scoring component. This test-retest pattern should be scored by an expert scorer to determine if the instrument, without interference by the list of aesthetic judgment premises, will produce similar category placement in two administrations over a three-week or longer period of time.

2. A series of postcard reproductions similar to those utilized in the UNESCO study (2) could become the basis for a study to determine predictive validity and stability for this instrument. In this research design, postcard reproductions would be used to confirm the selection of an aesthetic judgment premise by a student or an expert scorer. For example, a student who is placed or places himself in the mimetic category would be expected to select the most realistic of a series of paintings as the best work and place the most abstract or non-objective on the lowest end of a value continuum. In contrast, a subject expressing a judgment in the organistic category (primarily formal) would be
expected to select a Mondrian or an Albers in preference to a Rubens or a DeKooning. If the hypotheses are accepted, validity in terms of predicting future behavior toward an object could be determined. Producing a choice of an aesthetic premise matched by a choice of an art object on that same premise would provide evidence of stability.

3. One of the questions raised in the discussion of the hypotheses was the possibility that the category choice from the self-scoring component was a more accurate self-report by the student than the written statement procured in part one of the instrument. Administering the instrument using slide pairs and only the self-scoring component on a test-retest basis or in the research design described in the preceding paragraph may provide some answers to this question.

Questions and Observations

The review of related literature produced a number of research questions and relevant observations.

1. The decades of relying on choices by a panel of experts to establish validity for an instrument has precluded serious investigation of the impact of directed encounters with art works on subsequent aesthetic judgments. As a probable result, the literature related to this area of research is almost totally theoretical. The assumption that aesthetic judgments are affected by training, however,
permeates the total structure of art curriculum in both studio and non-studio areas.

2. This assumption of the relationship of training to sophistication of judgment was confirmed by Wilson's 1972 study of "The Relationship Between Years of Art Training and the Use of Aesthetic Judgmental Criteria by High School Students" (4) as well as in the research results reported by Battorf in 1947 (1). However, William's research (3) which used colored cardboard blanks produced no significant variation in scores by artists and non-artists, a direct contradiction of the Wilson and Battorf results which may be caused by the design of the study, the use of non-art stimuli, or the possibility that the degree of training has no relationship to aesthetic judgment.

3. Based on the review of the literature, available information regarding aesthetic judgment can be summarized as follows:

   a. Aesthetic judgments are made on a variety of bases, only some of which are considered adequate by aestheticians.

   b. Training in art may or may not affect the degree of adequacy of those judgments.

   c. Attempts to measure aesthetic judgments which have employed something other than full-color reproductions of art objects and directions to make a judgment regarding them are deemed low in content validity.
d. There appears to be a strong relationship between aesthetic appreciation and sophisticated aesthetic judgments.

4. The following questions can be raised as a result of the preceding observations and the related literature on which they are based:

a. What factors influence change in aesthetic behavior?

b. Is it possible that adequate aesthetic judgments result from something other than training in art?

c. What aesthetic behaviors and premises are employed by non-art trained adult learners?

d. What justification exists in the research literature for the art curricula in use or under development which is based on the assumption of the ability to affect aesthetic behavior?

e. What sociological, educational, experiential, demographic, chronological, or sexual factors impinge upon the aesthetic judgment premise employed by a subject?

f. What implications would these factors have in art curricula?

g. If an aesthetic premise can be determined, what are the most effective and efficient teaching strategies which may be employed to modify that premise?
h. If adequate judgments are based on sensitivity to the art object, how can that sensitivity be enhanced?

i. What impact does historical information or information on art production methods have on aesthetic judgment?

j. What implications will the results of the previous three questions have for the content and sequence of art appreciation courses?

5. Continuing investigation into the nature of aesthetic behavior, its components, and its basis for change, must be considered a critical research priority for art education.

Possible Applications

Among the possible applications of this aesthetic judgment instrument following the suggested additional studies of validity and reliability are as follows:

1. The instrument can be used to assess the entering and exiting aesthetic behavior of an individual student for an entire course or a unit of study. The instrument would make possible precise recording of change from one category to another.

2. The instrument can be used as a diagnostic tool to determine appropriate art activities to influence aesthetic behaviors. An example would be a student whose expressed
aesthetic premise is in contradiction to the type of art work he produces. A graduate student in this investigator's experience professed a deep commitment to art as communication with the masses yet consistently produced highly formal and intellectual art works. Use of the instrument clarified this incongruence and provided a basis for resolving the dilemma the student experienced. On a less philosophical question, the prescription might be simply to discover the fallacies in an aesthetic premise such as "it took so much time it has to be good" seen in other contexts.

3. The instrument can be used to determine the percentage of a population of students using least adequate, adequate, and most adequate judgment categories and the degree of change from one category classification to another as a result of a unit of study. The data obtained in this application would permit results indicating that "while fifty percent of the population used inadequate judgments at the outset of the semester, seventy-two percent of the population utilized the most adequate classification of judgments by the end of the course."

4. The instrument can be used to determine the relative effectiveness of varying teaching strategies on the aesthetic judgment making ability of the subjects.
Suggestions

As a result of this research effort, other investigators may profit from the following suggestions:

1. Because numerous students are susceptible to the notion that an authority figure knows best, particularly when confronted with unknowns such as art works, avoidance of any overt or covert expression of preference by the test administrator in the presentation of the slides is essential.

2. For the same reasons, use of masterpieces with the accompanying statement, "Both art works are recognized masterpieces," will avoid the "guess which one is really awful syndrome."

3. Further attempts to develop aesthetic instruments will probably proceed more efficiently if the guidelines developed for this instrument are adhered to closely. The critique of available aesthetic judgment instruments on which those guidelines were based eliminated a number of instrument methodologies from possible consideration as research tools.

4. Qualifications of an expert scorer should include thorough acquaintance with the Wilson categories as well as experience in deciphering the intent of a statement about art made by a non-art trained subject. The vocabulary of art is highly specialized and numerous phrases and words are
used by non-art subjects to approximate that vocabulary. Professionals who have not been actively engaged in the teaching of non-major students may not be adequately skilled in that type of translation. Further, until more research is conducted, any assumption that non-art trained students will make less adequate judgments should be avoided at all costs.

5. It is exceedingly important in the administration of the instrument to indicate that the subjects' personal opinions are the objects of the test. Concerns regarding spelling, handwriting, and "what answer do you want" should be alleviated as much as possible.


APPENDIX

PART I

This is not a test. It is a survey of opinions about Art works. There are no right or wrong answers. Please print carefully.

Name_________________________ Social Security No. ________

College _________________________

Have you ever taken or are you currently enrolled in:

Art Appreciation Yes ☐ No ☐
Music Appreciation Yes ☐ No ☐
Humanities Yes ☐ No ☐

Have you ever seen the TV Humanities series? Yes ☐ No ☐

The two Art works showing on the screen are both by well known and respected artists. Select the work you consider to be the best. Mark your answer below:

The best work is the one on the left. ☐
The best work is the one on the right ☐

In the space below, write your reasons for your choice. Please use complete sentences.

When you have finished your statements, raise your hand. The instructor or monitor will give you a second set of instructions.
PART II

Listed below are a number of reasons that people give for deciding one art work is better than another. Look over the statement you have just written. Then, from the groups of comments printed here, select the one group which seems closest to the reason you have given. You probably won't find an exact match in wording, but you should find statements with a close match in meaning. Choose only one group.

For example, if you have decided that the painting on the left is better because it looks more like the real thing, you would choose the group of statements labeled B.

When you have decided which group of statements is right for your reasons, put the letter for that group in the box in the right hand corner of the page on which your statement is written.

Turn all of these materials in to your instructor.

A] It is new and original. It hasn't been done before. The artist has made something different. Good art should be original.

B] It looks like a photograph. It is very realistic. It looks like the real thing. Good art should look like something.

C] Lots of people like this kind of art. It has been shown in magazines and books. Teachers have said it was a good work of art. It looks like an old master painting. It looks like the art works they put in museums.

D] It looks like the artist knew what he was doing. It has taken a lot of time. The painting is done well. Good art work should show that a lot of care has been taken with it.

E] I like it. It appeals to me and I know what I like.
All the colors blend nicely. Everything fits together. The shapes are interesting. All the parts of the work relate to the whole work. Good art work should have a feeling of unity and completeness.

The art work is inspiring. It portrays good ideas. Its message is acceptable. Good art work ought to be about good things.

It has everything a painting ought to have. Everything I think a work of art should have is there. My favorite colors are in the painting. It is my idea of a perfect painting.

I enjoy looking at this painting. It gives me a great deal of pleasure. I can imagine that I would enjoy looking at it for quite a while. Good art work ought to cause feelings of pleasure or pain.

The painting really communicates an idea. The message in the painting is clear. The painting says something important and says it well. Good art should express an idea.

The painting is forceful and strong. It has a great impact on me. It is very lively. The painting is very expressive of feelings. Good paintings ought to cause an intense and rich experience.

Nothing in any category fits my answer.
BIBLIOGRAPHY

Books


Bell, C., Art, New York, Frederick Stokes, Inc., 1913.


Articles


Reports


Publications of Learned Organizations


Unpublished Materials


