

Validating a Measure of Aesthetic Development for Museums and Schools

by Abigail Housen

originally published in *ILVS Review*, Vol. 2, No. 2, 1992

(Please note: This article refers to charts and tables which appeared in the text when the article was published. Due to technical difficulties, it has been impossible to embed them here. To receive a faxed copy of the charts, e-mail a request with your fax number, to info@vue.org. We apologize for the inconvenience.)

Stages of aesthetic development offer museums a means to assess their programs and audiences. This study examines the validity of Housen's model of aesthetic development. Second and fourth graders participated in a museum education program at Bard College over several years. The Aesthetic Development Interview (ADI) was used repeatedly over a two year period to assess program impact. The data confirmed the validity of the ADI in several ways. Elementary students could perform the ADI and all of their responses were codable. Early stages corresponded well to elementary school aesthetic judgments. The ADI discriminated between experimentals and controls, with high test-retest reliability, suggesting that the program produced stage change. The author concludes that the ADI can be useful in assessing the effectiveness of museum education with naive audiences.

The growth in museum attendance in the last twenty years has increased interest in understanding the aesthetic thinking of visitors and in methods for assessing the impact of educational programs for promoting aesthetic understanding. Most investigations have been descriptive, focusing on the presence or absence of a particular variable or category. There is much information about visitors' length of visit, educational level, and exhibition preferences, but these discrete and fragmentary findings add little to understand how visitors perceive or think about art or their art museum experiences (Housen, 1979; 1989).

Stage theory offers a theoretical framework for explaining the characteristic way a person reasons. The author extends this framework to viewing experiences to describe and measure characteristic ways visitors experience art (Housen, 1979). These ways progress from simple describing and viewing art. Stages affect a visitor's movement patterns, label reading, vocabulary and other ways of reacting to art. Knowledge of the stage of a visitor's aesthetic development, in principle, can predict how visitors will approach and describe a work of art. Potentially, stage theory may provide a useful theoretical framework for evaluating and reliably predicting the impact of art education material and museum programming. Developmental psychology has mapped the evolution of thought and judgment in several domains: Piaget in mathematical concepts Kohlberg (1981) in moral judgment, and Loevinger and Wessler (1976) in ego development. Lowenfeld (1957) mapped the development of children's art-making abilities. But, it was not until the '70s that researchers began to use a developmental model for exploring the evolution of the thinking and the viewing of art. Researchers attempted to develop stage models for the growth of thinking about art objects (Brunner, 1975; Clayton, 1974; Coffey, 1968; Gardner & Gardner, 1970, 1973; Gardner, Winner, & Kirchner 1975, Parsons, Johnston & Durham, 1978, 1987). Loevinger and Wessler's

psychometric procedure (1976), summarized by Housen (1983a), is neo-Piagetian. Its basic approach is to sample a large number of open-ended responses and then use these to create a distribution of scores that can be matched to typical distributions that empirically define each stage. However, the method differs from Piaget's in that the stage are not defined solely by a set of logical distinctions. Instead, stages are defined empirically by the response and thought patterns found to cluster together.

Since the late 1970s, the author has applied the stage model to study the aesthetic reactions of samples of adolescents and adults (ages 14-80) to two-dimensional works of art. These samples varied in socioeconomic level, ethnicity, and educational levels. One outcome of these studies was the identification of five stages of aesthetic development. The measurement methodology was based on the use of "stream-of-consciousness" interviews from which a score is derived representing where an individual "fits" into five stages of aesthetic development.

Few art of museum education programs have employed developmental measures as a way to assess their impact, or as a basis for designing educational materials. One obstacle has been the lack of reliable evidence to validate the existence of aesthetic development stages. An opportunity to determine the validity of the author's aesthetic development measure (Housen, 1983a) became possible when the Education Department at the Edith M. Blum Art Institute at Bard College asked the author to evaluate an art education program in progress involving 80 children over a five-year period.¹ The results from this effort are reported in this paper.

Stages of Aesthetic Development

Earlier applications of the stage model with adolescents and adults (Housen, 1979) identified five primary stages of aesthetic development. Except for one stage, one transitional stage was found between each primary stage.² The primary stages may be described as follows:

I. Accountive Stage. Viewers are storytellers, using concrete observations, personal associations and senses to create a narrative. Their evaluations of the art are based on what they like and what they may know or think they know about the art. Comments are colored by emotional terminology as viewers seem to enter the work of art becoming part of an unfolding drama.

II. Constructive Stage. Viewers build a framework for looking at works of art, using their own perceptions, knowledge of the natural world, and social, moral values, and conventional world views. If the work does not look the way it is "supposed to" (for example, a tree may be orange instead of brown or themes of motherhood may be transposed into wars on sexuality), then the viewer judges the work "weird," lacking or of no value. The craft, skill, technique, hard work, utility, of function are not evident. Emotional responses disappear as viewers distance themselves from the work of art, focusing on the artist's intentions.

¹ The Bard program was funded by the Arts in Education Program of the New York State Council on the Arts, the National Endowment for the Arts, the Edith C. Blum Art Institute at Bard College, and the Andy Warhol Fund.

² There are two transitional stages between Stages II and III.

III. Classifying Stage. Viewers describe the work in analytical and critical terminology similar to art historians. They identify the work as a place, school, style, time, and provenance. They decode the surface for clues, using their library of facts and figures. Once categorized, the viewer explains or rationalizes the work's meaning and message.

IV. Interpretive Stage. Viewers attempt to create some kind of personal encounter with a work of art. They explore the canvas, letting possible interpretations of the work slowly unfold; they point out subtleties of line, shape, and color. Feelings and intuitions take precedence over critical skills, as these viewers allow the meaning and symbols of the work to emerge. Each new encounter with a work of art elicits new comparisons, insights, and experiences. Viewers accept the idea that the art's identity and value are subject to re-interpretation, and see a given interpretation subject to change.

V. Re-Creative Stage. Viewers, having established a long history of viewing and reflecting about works of art, now "willingly suspend disbelief." A familiar painting is like an old friend known intimately, yet full of surprise, needing attention on the daily level as well as on an elevated plane. In all important friendships, time is a key ingredient. Knowing the ecology of the work—its time, its history, its questions, its travels, its intricacies—and drawing on their own history with the work, in particular, and with viewing in general, allows this viewer to combine a more personal contemplation with one which more broadly encompasses universal concerns. Here memory infuses the landscape of the painting, intricately combining the personal and the universal.

Since the mid 1980s, the ADI, the master Scoring Manual, and stage descriptions have been employed by several museums to analyze demographic data as well as for assessing the impact of art education programs on aesthetic development (Housen 1987, Housen, Miller & Yenawine, 1991).

The Red Hook Art Education Program.

The collaborative arts in education program was initiated by the Edith C. Blum Art Institute at Bard College with a local elementary school, the Mill Road School in Red Hook, N.Y.³ The arts in education program centered around exhibitions at the Art Institute which ranged from classical Greek and Roman art to 19th century paintings of the Hudson River School to experimental works in photography and video. One year the program involved the local history component of the school curriculum, *viz.* "Charmed Places: Hudson River Artists and Three Houses, Studios, and Vistas." Preparatory teacher activities included one-day workshops, pre-visit meetings with parents and students by grade level, and pre-visit teacher packets (vocabulary lists, timelines, maps, biographies, activities).

The program centered around the museum visit itself. The visit to the museum lasted from 2 to 3 hours focusing on one gallery. Each museum visit included:

- a) Informal discussions of the exhibition, background information, and comparisons to previously seen exhibitions.

³ Funded by the New York State Council on the Arts.

b) Use of student activity sheets (by skill level) that encouraged active looking and independent viewing.

c) Return to the exhibition gallery in small groups or as one large group (depending on size and grade), using the activity sheets for review. Students could ask questions, discuss artworks, and express opinions. An artist-in residence was often available to work with students during this 2nd visit.

d) Hands-on production of art by students based on gallery paintings. Activities included a collage for “Haley’s Comet”; landscape dioramas for “Charmed Places”; jewelry from found objects in “The Arts at Black Mountain College”; and pseudo-archaeological digs for the “Herakles” exhibition.

e) A group discussion of the exhibition and overall visit, focusing on critical thinking skills: *Why were you here? What did you learn? How does this exhibition compare with exhibition you have viewed at the Art Institute before? Which is your favorite artwork? Why?*

After the museum visit, in-school follow-up activities took place, such as writing, storytelling, art projects, plays, and music. Reading resource material also was available for interested students. In addition, a full day Spring celebration was conducted outdoors where students and teachers visited “learning stations” based on the previous museum visit (Egenberger, 1991). This collaborative arts in education program had been going on for two years before the author was contacted to assess the program.

Design, Subjects, and Schools

Two independent groups (n= 40 each) of 2nd and 4th grade students were administered the Aesthetic Development Interview (ADI). One group (AE: experimental) participated fully in the arts in education program at Red Hook School. The second group (C: control), of comparable size at a nearby school, was not in the program. The 40 students in each group were selected at random from each school. Both demographic and museum biography questionnaires focused on student background and interests. Both groups were similar in age and socioeconomic status (predominantly white, working and middle class).

Measuring Aesthetic Development

The Aesthetic Development Measure described here is discussed in terms of (1) the Aesthetic Development Interview (ADI) and (2) the translation of interview data to aesthetic stages. This measure was used to assess student progress in the two groups described above for two consecutive years with on-going longitudinal assessment planned.

Aesthetic Development Interview (ADI)

The ADI is designed to identify aesthetic development stages, averaging between 10 and 20 minutes beginning with an open-ended query, such as *What do you see? What are you looking at?* Students are asked to talk aloud in stream-of-consciousness fashion about

what he/she “saw” in a reproduction of a given two-dimensional work of art until the student had no more to say. Because the respondents freely talk aloud in stream-of-consciousness fashion about a work of art, they are not prompted by specific questions and are encouraged to express perceptions and thoughts as they occur. The only prompts given during the interview are: *What are you looking at?* and *Is there anything more?* Interview data are tape-recorded and transcribed.

Independent Thought Units (TUs)

Transcriptions of interview data are broken down into independent thought units (TUs). TUs are complete ideas varying in length from short phrases to longer sentences.

Examples: I see people;
The kid's taking off his shoes and the girl's;
I like red because it's so bright.

The number of TUs varies widely from one respondent to another. To provide a fixed number of TUs, 15 TUs are selected at random from the total. These 15 TUs become the TU sample to be analyzed and scored for each student.

Translating TUs to Aesthetic Stages

The “meaning” of each TU is next translated into an appropriate stage category. There are 14 domains and 70 sub-categories each of which defines a different genre of remark. Domains are distinct cognitive or affective expressions by individual TUs, such as associations, observations, preferences, evaluations, comparisons, assertions or types of thinking. Each TU is assigned to one domain and sub-category which, in turn, can be translated into a stage score.

The eventual aesthetic stage for each student is determined independently based on two procedures:

1. *Empirical Rating*: Using the scoring manual, each of the 15 TU samples from each respondent is coded to a best-matched category. These 15 TUs yield a distribution of stage scores for each individual. This set of scores is then collapsed to a single stage score for each subject. By matching each individual's distribution to an empirically derived “master curve” for each stage, one can determine which stage curve the individual best matches.

2. *Clinical Rating*: An independent trained rater examines the transcription of each interview and gives a “clinical rating of the subject's stage score.

These ratings are compared and a final aesthetic stage score is assigned to the subject, representing the best match of the empirical and clinical ratings. This method yields a high inter-rater reliability (Cone's Kappa averaging 0.90) for coding the TUs and for the final stage assignments (Housen, 1983a). Finally, a *cumulative stage index* is created by weighing scores at each stage and cumulating these across the 15 TUs. This converts the individual distribution into a *continuous variable*.

A summary of the ADI scoring procedures is as follows:

1. **Identifying thought units.** The transcribed interview is parsed into thought units (TUs), each containing a complete thought. “that lady looks kind of funny standing there/ ...I’m looking at the girl doing whatever to her hair.../ ...I like the bathing suits.../ ...The sea... I don’t know if the sea looks like that... / ...This could be a boat...”
2. **Selection of thought units.** Fifteen thought units are selected from each interview. Each transcript is divided into thirds, and five TU’s are sampled from the beginning, middle and end of the transcript.
3. **Matching thought units (TUs).** The Scoring Manual (Housen, 1989b) combines 14 Domains and 70 Sub-categories against which TUs are matched. Examples of coding for specific thought units are as follows: (I) (IV), etc. = name of Domain; subscript = Sub-Category)

<u>Domain</u>	<u>Sub-Category Description</u>	<u>Most Frequent Stage</u>
<i>That lady looks kind of funny standing there.</i> Evaluation (IV)	Qualitative judgment is based on personal criteria about what is real.	Stage I/II
<i>I’m looking at the girl doing whatever to her hair...</i> Observation (I)	Animation	Stage II
<i>I like the bathing suits.</i> II. preference (II)	Likes/dislikes referring to objects in work	Stage I
<i>The sea...I don’t know if the sea looks like that.</i> Assertion (VII)	Personal statements which are incontestable	Stage II
<i>This could be a boat.</i> Association (III)	Random personal association	Stage II

Note that each category has an associated stage. This is empirically determined from the population which uttered this class of remarks. The category’s Stage score” represents the most frequent stage of subjects who most regularly use this category. To obtain a high level of inter-rater reliability, coding must be done only by coders who have been trained and certified in a workshop.

4. **Assigning a stage score form each thought unit.** The score assigned to each TU is the stage score of the category to which it was matched. Each TU is rated independently, divided from other elements of the interview.
5. **Calculating TU scores.**
Individual distribution.

The number of thought units scoring at a particular level are counted.

<u>Stages:</u>	<u>I</u>	<u>I/II</u>	<u>II</u>	<u>II/III/IV</u>	<u>III</u>	<u>III/IV</u>	<u>IV</u>	<u>IV/IV</u>	<u>V</u>
Individual number of TUs at each stage	4	1	8	2	0	0	0	0	0

Cumulative distribution.

Individual scores are cumulated by stage to create a cumulative distribution.

<u>I</u>	<u>I/II</u>	<u>II</u>	<u>II/III/IV</u>	<u>III</u>	<u>III/IV</u>	<u>IV</u>	<u>IV/IV</u>	<u>V</u>
4	5	13	15	15	15	15	15	15

6. **Obtaining an overall stage.** This is a matter of matching individual distribution to stage classification rules. (Scoring Manual, Housen, 1983b), which have been empirically derived from the original sample on which the scoring system was developed. Classification rules are descriptions of Master distribution curves for subjects (from original sample) that have been clinically classified at each stage.

For this example, some useful relevant classification rules would be:

<u>Stage Classification</u>	<u>Rules for Classifying Stage</u>
Stage I	At least 5 ratings at Stage I
Stage I/II	At least a6 ratings at Stage I/II
Stage II	At least 8 ratings at Stage II
Stage II/III	Not more than 10 ratings at Stage II and at least 3 ratings at Stage II/III, or III

By using these rules, the distribution in this example would be rated as a Stage II.

7. **Clinical Rating of the Whole Interview.** A trained clinical rater reads the whole interview and assigns a clinical rating. (In this example, imagine a clinical rating of Stage II). Again, only certified clinical coders can accurately make these assessments.
8. **Comparison of Empirical and Clinical Ratings.** When two ratings match, the stage score for the subject is finalized.
9. **Creation of the Cumulative Stage Index.** The individual distribution is converted into a continuous variable by weighing scores at each stage and cumulating across the 15 thought units.

<u>Stage</u>	<u>I</u>	<u>I/II</u>	<u>II</u>	<u>II/III</u>	<u>III...</u>
Thought units	4	1	8	2	0
<u>Weight</u>	<u>x1</u>	<u>x2</u>	<u>x3</u>	<u>x4</u>	<u>x5</u>
Weighted Total	4	2	24	8	0
Cumulative Score	e = 38				

Assessment Phases

The Bard art education program took place over a four-year period, as shown in Table 1. Assessment of this program began in the third year, with follow-up in the fourth year. Thus, Group AE (experimental) had participated in the Red Hook program for two years and had experienced two museum visits before the author joined the project to assess student progress. As shown in Table 1, 2nd and 4th graders were tested before and after their third visit to the museum. Post-tests examined the degree to which the students absorbed critical information from the museum tour. Responses to the open-ended questions about the content of the tour were coded as either acceptable or unacceptable. There was also a follow-up phase with ADIs beginning in the Spring of 1989 for 3rd and 5th graders.

Post-test ADIs were administered in late Spring, 1988, after students were interviewed with a second reproduction from the same artworks. After the ADI and tour, students were then administered a second set of five open-ended questions (coded acceptable / unacceptable) focusing on the content of the tour, as follows:

1. What kind of a painting is this?
2. When was it painted?
3. How did the artist make it look the way it does?
4. Where was it painted?
5. How does it make you feel?

These questions were a more conventional way of assessing what they had learned from their museum experience. Experimental and control groups were compared in terms of correct answers.

Table 2 summarizes the chi-squares for these differences for each question for the 2nd grade experimentals on the second, fourth, and fifth questions and for the first and fourth questions for the 4th graders. The 4th graders did not respond to the 2nd question on *when* or the fifth question on "feeling".

Results

Selections from Interviews

Each student discussed a reproduction selected from the works of Edgar Degas, Winslow Homer, Thomas Hart Benton, Robert Havell, and Frederick Church. Works by Havell and Church were on exhibit at the Blum Gallery when the experimental students visited "Charmed Places: Hudson River Artists and Their Houses, Studios, and Vistas." The following selections from these interviews illustrate some qualitative differences between AE and C groups. For example, a few of the AE and C responses to What do you see? What are you looking at? were: (C = Control subject; AE = Experimental subject.)

C (2nd grade):

I see people. Well, the kid's taking off his shoes and the girl's...The dog's barking at the hole...The boat...the mountains...the birds...the sky...water...sand...shells...dog...people.

These responses express concrete observations along with simple tabulations that enumerate items in the artwork.

AE (2nd grade):

Well, it looks like the woods and it has a big house in the background. It has a lot of land. This is an artist's house, right? So he can probably look at the land and see what he sees and he could paint and it looks like he's got a lot of windows. I like the boats. I like that big mountain there and how the sky looks, and then you can see some of the hills in back and the water and there's so many animals. I have a lot of animals at my house and I like animals.

These remarks are richer and include observations, preferences, associations, and interpretations. The viewer is more aware of the artistic process and craft.

AE (2nd grade):

Well the dog is, like, barking at the boy. And the woman's, like drying off her dress and wiping her hair and ummmm, the other guys taking off his shoes so he can go swimming. The other guy, I think, he's ready to go swimming. I don't know if that's right or...and the birds they're flying in the air like turning and stuff doing tricks and stuff in the air.

The use of the active verb suggests that the ongoing event is occurring before the viewer's eyes. This is similar to the response given by the 2nd grade control at the beginning of his list. This characteristic, where the viewer unconsciously animates the work of art, is called "pre-semblancing" (Housen, 1983a). However, in this example, the child is beginning to become aware of her own responses and seems to be distancing herself from the artwork.

AE (2nd grade):

I like the farm because of its crops and I like crops because it helps people save lives because some people don't have food and they eat the crops...etc.

Note the beginning of the student's interest in deciphering the message of the artist.

AE (2nd grade):

I like the picture and I like the way it is like. I like the way it's like painted and I like the waves.

The child in this case is able to state preferences which he supports with observations about technique.

AE (2nd grade):

The paint the artist used, the colors, they're so nice. It's like you can walk right into it as if you were dancing...ummm. There's a pine tree. And a few raindrop shaped leaves...

The above observations and evaluations are personal responses. The viewer balances knowledge about intentionality with an ability to enter into the artwork, even responding in kind with a metaphor of her own.

AE (2nd grade):

Ummm, I like how he drew...Whoever made this painting...ummm, made this wire, cause it looks like it's real...cause my mother brought some wire from where she works.

Note the interest in and knowledge about artist, craft, technique, and skill, characteristic of Stage II.

AE (2nd grade):

Well, it's kind of beautiful...they're dancing like ballerinas. Their costumes are gorgeous. Like the colors in it and the scenery is perfect for a painting. It looks like they're dancing in Autumn - Fall. The leaves are red, orange, green, different colors and there's a rock in the background. Pretty.

This response combines evaluations, comparisons, preferences, and classifications. The viewer animates the reproduction, entering into its world, responding to its aesthetic features.

C (4th grade):

Well it looks like those kids fell in the water or something. And the girl's wringing her dress. There's a dog watching...There's a boat in the water. Sea gulls up there. There's three people on the other side. And that's it.

As the second-grade control and experimental subjects did, this viewer animates the work of art. Here the enumeration of the content is lengthier and more detailed.

C (4th grade):

What I see in this picture I see mansions, sailboats, a huge cliff...Rocks... the Hudson River...or a sea or whatever it is... Who knows? Who cares even? Oh boy! I see cattle grazing in the field...Cattle walking down to the place where people live... Cattle walking up... People selling stuff...A woman or a girl taking milk down... Trees... I see a huge mansion way up in the distance... I see a big ocean liner or something... I see a saddle... A dog... A greyhound, perhaps, or a husky, or a German shepherd...Who knows? Who cares? I see a stone wall...with a big house n the other side... and one big mansion...And I might say far off in the distance Frederick Church's house, but who knows? Who cares? Who wants to know? I see mountains way in the distance and a big cliff... Whatever it is... Who knows? Who cares? Who wants to know?

This viewer's cataloging of the content is lengthier still and more detailed, yet lacks an awareness about the artist's intentions. The concept of intentionality is yet not present. In contrast, note the following voice of a 4th grade experimental.

AE (4th grade):

And the way he painted it, it really looks nice...because, like the flowers and then he added, like detail, like making this, like little animal on the stone and moss on the stone... and it really looks like a nice painting. He...looks like that he was good using his hand...And the say he did the tree almost looks like it's real... And like the fields, they're probably plowing it so that they can plant their crops there... Have their animals feed on it... I like it, like how he made it so it really looks like over here...How it looks like the field is really behind the tree instead of right next to it... And here under the tree. The way he draws makes you, like think, about how everything was long ago. And like he drew... He added a lot of detail. he put flowers right by the tree... And then he made vines with flowers on them... Like you know most people just draw with solid lines, and he drew it so really well... It looks like a leaf.

These observations, rich in detail, refer to the artist's technique an skill. The viewer makes comparisons, associations and classifications from formal categories as well as from everyday life to support his preferences an evaluations. He both enters the world of the artwork, animates the work, and maintains a self-conscious distance from that world, aware of issues of intentionality.

AE (4th grade):

It's very realistic. The artist never really left any details out of it. It's almost real. And it shows the flowers and the bugs and the hay and even the background is good. Well, it's well drawn and it's almost... The clouds look kind of tired and they're hard working and they're not very rich. he even put the light in the shades... And the light is just really... You can really tell about the light because some trees are really dark and some are really light green.

This viewer's concern with detail, technique, and reality, are characteristics of Stage II. The viewer provides precise details, combining interest in formal details as well as in expression and meaning.

AE (4th grade):

The tree looks like an alligator's skin.

this subject provides an example of a response in kind an expressive metaphor.

AE (4th grade):

It looks like people are dancing and they're having a good time. To me, when I look at the picture, it makes me feel like everything's really graceful. The colors are, some of the colors are really weird... The greenish colors in the background... The ballerina is quite pretty. I think it's a very pretty attractive background. At first, when I was it, I thought it was a big ranch. I like how he drew the trees, it gives it detail, because you see the church off in the distance. You're right there, behind the trees. It makes me feel like I'm peeking over to see what's happening back there.

Category Adequacy

There were 80 subjects both pre- and post-interviews which accumulated to 1,200 TUs. Virtually all responses could be matched to the existing (Master) scoring categories. Thus, the categories accommodated the broad range of responses of children 8-12 years of age, despite the fact that these subjects were younger than the sample from which the Master categories were derived. However, the frequent use of one category (I) suggests that subdivisions might be appropriate for future studies.

Model Accuracy

Figure 1 shows combined data for all subjects for the initial half-stages of the aesthetic development scale (Stages I to II/III). The exception that these young subjects should test at the early stages of aesthetic development was confirmed. Their scores were clearly skewed to the lowest stages and concentrated at Stage I and I/II. All 80 subject's protocols could be classified in these initial stages, in terms of both empirical ratings and the clinical stage ratings conforming to what one would reasonably expect in terms of stage distributions for younger groups.

Aesthetic Stage Changes Over Time

Comparing the distributions of experimentals and controls confirms what appears to be true in the protocols: that the two groups differed in aesthetic level in the two years *prior* to the beginning of the assessment period. Specifically, percentages for experimentals at Stages I, I/II, and II are nearly equal, while there are three times as many control subjects at Stage I as at Stage I/II. A cross-tabulation of these stage scores yielded statistically significant differences (chi-square = 9.014, $p < .03$).

Figure 2 shows the stage distribution by grade. The difference between experimentals and controls is much larger for the older (4th grade) students. Second graders are concentrated at Stages I and I/II for both controls and experimentals. The picture is different for 4th graders. Controls are mostly Stage I; experimentals are mostly Stage II.

Short vs. longer term changes. Figure 3 compares the pre- and post-stage scores for experimentals and controls and shows only a negligible shift during the assessment year observation period. Results are consistent with the original expectation that changes in aesthetic stage would occur over *longer* intervals, not necessarily over short intervals of a few weeks or months. Experimentals did not show Stage shifts ever the 3-month period of the assessment year. The tests of assessment year differences between pre- and post-stage scores were not statistically significant. Cumulative pre- and post-scores provide a more sensitive analysis of score shifts. However, a one-way analysis of variance of experimental vs. control groups failed to show significant differences in aesthetic stage gain.

While stage change appears negligible in the short run for both experimentals and controls, there were sharp differences between experimentals and controls in long-term change in stage scores, as shown in Figure 4, which presents composite changes in Stage scores from younger students (2nd grade) to older students (grade 5). This suggests a kind of “synthetic view” of change that might plausibly take place in a single group from the 2nd to 5th grade. The combined data in Figure 4 shows mean stage scores for each group over the 4 year period. The experimental group shows long-term change in Stage scores while there was little, if any, change for controls.

Test-Retest Reliability

The stability of short-term test scores offers a kind of “back-handed evidence” of test-retest reliability. Test-retest studies are seldom done in the context of an intervention to promote change. Yet the absence of change during assessment year ADIs allows an unusual examination of test-retest reliability. Do individuals retested over a relatively short interval show stable scores that are highly correlated? The effects of the Bard program aside, one would expect individuals tested at two points in time to exhibit the same stage score. Of 54 subjects in this sample, fewer than 20% show a shift in scores, and those only half-stage shifts. Furthermore, increases roughly balance decreases, to suggest that even those shifts are the result of measurement errors. The stability of scores equivalent for both experimentals and controls. A correlation for the assessment year cumulative scores on pre- and post-tests is also high ($r+.938$) and is another index of test-retest reliability.

Stage Onset

The cumulative score also throws light on the range of ADI responses that differentiates a “beginner” from an “advanced” Stage I viewer, using cumulative stage scores of subjects rated at the same stage. Based on *relative* frequency, 24 sub-categories of TUs were organized into five categories of responses to the artworks, as follows:

- A) Descriptions: The viewer describes what he sees.
- B) Animation: The viewer energized the work with life-like characteristics.
- C) Personal Reality: The viewer expresses a personal or idiosyncratic interpretation.

- D) Formal Evaluation: The viewer judges the formal properties of the work such as color, shape, and line.
- E) Assertion: The viewer makes unqualified declarations.

The “beginner” (A and B) and more “advanced” categories (C, D, and E) do not appear with equal frequency. Figure 5 shows how these response categories were distributed as a function of stage scores. Note the different categories were distributed as a function of stage scores. Note the different thresholds at which response categories A through D emerge and the marked differences in the maximum frequencies of the response categories between stage score levels. For example, lower stage scores of 20-30 show a very low frequency of “personal reality” responses (C) and sharp increases in the frequency of type C responses at higher stage scores. There appears to be a hierarchical order from “beginner” (A) to “advanced” (D) responses.

Table 3 compares the distributions of these response categories of Control and AE groups at different stage score levels. Note that few Controls displayed “advanced” response categories (only 15% exhibited C and D categories). At any given stage, few Control subjects displayed “advanced” response categories. The Experimentals showed a different pattern. At least half displayed “advanced” responses such as C, D, or E (level 3). For a given stage, there was an independent tendency to display “advanced” categories and this tendency appeared to be facilitated by the Bard art education program.

While there was no short-term gain in aesthetic stage, the ADI up short-term shift in overall response category levels. While this shift was not statistically significant ($p = <.06$) for the AE group, the shift was in the expected direction; i.e. there was not change for the Controls from A and B response category levels, while the AE group moved from A and B responses at the beginning of the program to a C or higher level of responses at the end. In terms of up / down shifts in A and B to C and D response categories, while not statistically significant, fewer 2nd grade Experimentals began the pre-assessment period at “ceiling levels”; more 2nd graders in the Bard program showed category gain (chi-square = 5.02, $p = <.08$).

Discussion

The observation offered by the Bard program support the validity of the aesthetic development measure in a number of ways. First, children as young as those in the 2nd grade were able to respond to the ADI. Second, the existing scoring manual was capable of classifying all responses of an elementary school audience. Third, children’s responses reveal, in a detailed, yet natural way, a range of influences of the program on their thinking. Fourth, the initial stages at which all students tested, Stages I to II/III appear to accurately characterize the range of thinking of students about their art experiences in the Bard program.

Construct validity is not established in any single observation, nor in any single study. Instead, one looks for patterns in the data which match the tenets of the aesthetic development model. What is noteworthy is that from the wide variety of observations, each, in some sense independent of the others, confirms the predictions which flow from the aesthetic development stage model and its method of measurement.

Inter-rater Reliability

The Bard program required the repeated testing of students, especially experimentals, over a multi-year period. Assessing test-retest reliability was not the focus of the Bard art education program, but the assessment of whether the program had an impact on or accelerated the aesthetic development of the young students. In fact, if aesthetic development occurred during the program, this would *reduce* the apparent test-retest reliability; i.e. experimentals would *not* show the same stage in the next reliability, this would mean that little growth had been stimulated.

The key to resolving this conflict appears to lie in the interval of time one chooses to compare the initial test and the retest. Since ADI assessment occurred in the assessment year as well as the follow-up year, there are many pairs of tests one could correlate.

Under the circumstances, it appeared worthwhile to assess test-retest reliability by comparing scores taken as close as possible in time. The pre-post assessment administered during the assessment year involved a rather short interval between ADIs. Of the 80 students tested over this year, only 3 students showed a change in ADI score (up or down), and these 3 showed only a half a stage change. Combining the experimentals and controls and correlating the cumulative scores for each subject, the resulting test-retest association was extremely high ($r = +.938$) which might be considered strong evidence that the ADI and coding methods were producing stable and consistent results for a given subject with repeated measures over the period.

If the Bard program was in fact accelerating aesthetic development, was the ADI capable of tracing these effects? Again, the interval between the ADIs appears to be the key. When pre- and post-scores in the assessment year were compared, both experimentals and controls remained the same – no shifts, no gain for either group.

However, if one compares AE and controls in terms of time between ADI administrations, the patterns between pre-assessment and the subsequent two years are different. This could be due to differences in the backgrounds of the experimentals and controls rather than the Bard art education program. However, comparing the backgrounds from a demographic questionnaire given to both groups show no significant differences in backgrounds or other demographics or attitudes. If so, the Bard program may be responsible for differences in ADI scores. These similarities in backgrounds of AE and C groups also continued in the follow-up year, in spite of the fact that the C group in the follow-up year for these controls was similar in background and aesthetic development to C group in the assessment year.

These findings have a double significance. First they suggest that the Bard program accelerated the aesthetic development in children as young as elementary school levels. If so, this extends our understanding of aesthetic development, revealing previously unknown patterns about the timing and extent of aesthetic development at these early ages. From the point of view of construct validity, the results also suggest that ADI is capable of evaluating the impact of the Bard program on an aesthetic stage – i.e., to differences in the ways children “think” or express themselves about art. Beyond overall stage scores, the aesthetic interview method seems to have successfully elicited and indexed shifts in aesthetic stage development. While these shifts fall short of an overall stage change, apparently they can be predictors of later stage changes (see discussion below). Taken together, the variety of the evidence argues for construct validity of the aesthetic stage model and the measures used.

Dynamics of Change

The data also provides the first glimpse of an answer to some fundamental questions: can aesthetic development be accelerated, and if so, at what rate? Can an intervention at the second grade produce accelerated growth? While the Bard data suggests that program for elementary art students can produce accelerated growth, the data also reveal much that was previously unknown about the dynamics of such changes.

For example, what period is required to obtain a stage change? Certainly the data suggest that two years are adequate. Indeed, the comparison between the assessment year post-scores and follow-up year post-scores show that one year is sufficient to obtain an observable stage change (Figure 4). This shift occurred in a population which had not show stage change only a few months before (during the assessment year), but had shown change at the category level. Over the short run periods, apparently a stage change should not be expected even among experimentals. Rather, one must analyze at the category level to see statistically significant changes.

Data from the follow-up year show that after additional observation, stage differences between experimental and controls continue accelerate (Figure 4). The fact that the *post*-tests of 3rd grade were on the same growth trajectory as 4th grade *pre*-test measures, suggesting that the program alone may account for the high stage scores of the 4th grade experimentals during the assessment year.

Conceptual Thresholds. Findings on the frequency of category clusters (A, B, C, D, E, p. 229) suggest that expressing and thinking about art experiences that define aesthetic stage levels may first require the emergence of new concepts. The existence of such thresholds, in turn, may shed light on why experimentals and controls were comparable on the two demographic questionnaire items, technique and feelings (Table 2). The experiencing of art objects may be influenced by perspectives about art that had not been developed until the end of Stage I and/or the beginning of Stage II. Most 2nd graders in this sample are at Stage I. Perspective, a rather abstract concept, may still elude 2nd and 4th grade beginners. Also, the characteristic of Stage II viewers to avoid discussing their feelings, would account for the reticence of Stage II 4th grade experimentals to exhibit Stage II TUs during the ADI.

Incubation. The clear gains in aesthetic stage by the 4th grade experimentals, noted even before their first assessment year museum tour, suggest that they learned, retained, and assimilated information from prior museum visits during pre-assessment years. Yet, the students did not or could not incorporate *any* of the content information from this tour in their post-test ADIs. This was in spite of their ability to answer content questions about the tour. Taken together, this raises interesting questions about aesthetic growth. Effective assimilation of aesthetic content from art experiences may depend on more than just the ability to answer short-term memory questions. The ability to utilize experiences with art is a lengthy process and demands time. The slower rate of changes in stage scores over a short period, when combined with the significant shift in stage scores over a year, suggests that these students required an incubation period to assimilate and apply the aesthetic information. Short-term knowledge gain may not produce observable change in aesthetic stage because retention of short-term knowledge may not assure the application of this knowledge.

Developmental Ceilings. One puzzling question in the Bard data was why there was so little development beyond Stage II/III. Almost 20% of the 5th grade students

were at Stage II/III. However, no student made the transition to Stage III. This suggests that there may be prerequisite for aesthetic change that were not present in the Bard program. It is plausible that the ability to reason abstractly, the Piagetian stage of formal operations, is first required for movement to Aesthetic Stage III (Classifying Stage). While some 5th grade students appeared ready for Stage III (i.e., they could observe, categorize, classify, discriminate, and compare aspects within a work of art), most were, at best, just on the cusp of formal operations. Previous studies (Housen, 1983a; Housen, Miller & Yenawine, 1991) have reported that even adolescents are predominantly at Stages I and II. It also is possible that the curriculum may not have contained ingredients necessary to facilitate Stage III thinking. The Bard program, was designed for elementary age students and may need to be augmented to encourage more “advanced” students to move to Stage III.

Evaluating Art Education Programs

This study helps to confirm the relevance of the aesthetic stage model to programs of art education, both inside schools and museums. The stage measure provides a potentially improved approach to evaluating art education programs. Further, the stages and the scoring manual point to way for refining program content so that it is more attuned to the developmental process.

The Bard data suggest that there indeed may be a need for a new approach to evaluating the impact of art education programs. First, what constitutes an appropriate index of aesthetic change? Is it best to focus on measuring attitude changes and the recall of factual materials? Or would it be more realistic to assess shifts in the aesthetic response itself—i.e., shifts in the overall pattern of perceiving and processing information about objects of art?

The Bard data not only suggests that shifts in aesthetic responses can be detected, but also that the ADI measure was sensitive enough to identify factors in their art education program that may have been responsible for such shifts. A broader construct like aesthetic development seems a more germane way to assess a program’s outcome than employing arbitrary lists of factual and attitudinal responses. Traditional evaluation methods (questionnaires, concept tests, attitude surveys) have limitations in describing the way students learn about art. In this study, students who showed they knew they answers to some of the content questions (Table 2) were unable to apply this information during their ADIs. To the author, it seems unlikely that true/false questions about complementary colors or attributions about “schools of art” reflect the rich and complex thinking of students that were exhibited in the ADIs during and after experiencing art objects. A student may not remember or be able to articulate the names of complementary colors, or details of art history after such experiences, but this may not say much about their ability to respond in other meaningful ways to a work of art. Students and programs that focus on traditional simplistic approaches to evaluating development in art. education may fare poorly in the long run. And learners who may be able to recall the names of complementary colors or details of art history, but cannot talk comprehensively about works of art, may also fare as poorly as the programs that encourage traditional approaches to teaching (and evaluating) art.

If aesthetic stages are used in evaluating art education programs, a reassessment of how long-term vs. short-term impact are to be evaluated needs to be done. According to this study, aesthetic growth requires time before new experiences and ideas introduced by a program can incubate and reorganize the way in which learners think about what they see moment to moment in a work of art. If so, art education program

evaluations based on short term impact may end up rating a program prematurely, missing its longer term impact.

Extension of Theory

The data presented here suggest extensions of the aesthetic stage model in a number of ways. Previously, there had been no longitudinal sample with repeated aesthetic measures. The data revealed much about the dynamics of change that have not been previously observed. For example, it shed light on the *pace* of change and revealed elements and order of shifts in thinking. Finally, the Bard program itself threw some light on some of the kinds of stimulation that can facilitate aesthetic growth.

Micro-changes in Category Shifts. Data on the onset of stages extends theory about aesthetic development measure in several ways. First, groups of categories appeared to shift in a significant way in the experimental groups over the short run (periods as brief as three months), even though the overall stage did not change. Thus, stage change does not necessarily occur as a massive shift in all categories, but rather as a gradual transition, with some categories of thinking emerging first, and others later. Short run shifts in critical categories appear to be predictive of stage change over longer intervals—a real extension of aesthetic stage theory.

The measure appears capable of discriminating as well as indexing such shifts, which suggests that the measure can assess program impact both at the “macro” level (stage changes) and at the “micro” level (category changes). If so, the category changes extend our understanding of the utility of this measure.

Implications for Art Education Program Design

By supporting the validity of the aesthetic stage model, the Bard data also raise questions about program design. If aesthetic stages are a real phenomenon that can be reliably measured and mapped, this kind of information should be applicable in the design of educational programs. If stage ranges are predictable at certain ages, along with categories of thinking, category thresholds and category transitions, this should be helpful in refining art education programs. Several museums already are actively engaged in exploring these possibilities (Housen, 1989; Housen, Miller & Yenawine, 1991).

Teaching to the thoughts that naturally occur during early aesthetic stages in children would necessitate reducing the amount of information given about artists, schools and style that are characteristic of Stage III thinking. Contrary to standard practice, “being true” to the interests of the naive viewer appears to accelerate growth.

At early ages, aesthetic growth may turn out bear little relation to later conceptual frameworks. Perhaps, however, the ability of a naive viewer to “enter a picture” and “play” with the spaces and figures inside, may be comparable to “old friends” label that seasoned viewers of art give to painting. Naive viewers may have as much to tell about the end state in aesthetic development as do traditional Stage III museum goers, who rely so heavily on classification, styles and schools. Given the knowledge categories that the youngest group absorbed during museum visits in this study, the methods of art educators who emphasize exposing students to distinct disciplines in the arts seem questionable.

Perhaps the ultimate wisdom of the Bard program, and the basis of its apparent impact in promoting the aesthetic response, was that it permitted children to experience and play out their naturally evolving responses to art. The Bard program demonstrated

another way to prepare youngsters to enter the world of art by encouraging them to say for themselves what they see in a work of art.

References

- Bruner, C. (1975). *Aesthetic judgment: Criteria used to evaluate representational art at different ages*. Unpublished doctoral dissertation, New York, NY: Library, Columbia University.
- Clayton, J.R. (1974). *An investigation into the developmental trends in aesthetics: A study of qualitative similarities and difference in young*. Unpublished doctoral dissertation. Library, University of Utah.
- Coffey, A.W. (1968). *A developmental study of aesthetic preferences for realistic and nonobjective paintings*. Unpublished doctoral dissertation. Library, University of Massachusetts.
- Egenberger, Catherine (1991). Personal communication.
- Gardner, H. and Gardner, J. (1970). Developmental trends in sensitivity to painting style and subject-matter. *Studies in Art Education*, 12, 11-16.
- Gardner, H. and Gardner, J. (1970). Developmental trends in sensitivity to form and subject. *Studies in Art Education*, 14, 52-56.
- Gardner, H., Winner, E. , and Kirchner, M. (1975). Children's conception of the arts. *Journal of Aesthetic Education*. 9 (3), 60-77.
- Housen, A. (1979). *A review of studies in aesthetic education*. Unpublished paper. Cambridge, MA: Harvard Graduate School of Education.
- Housen, A. (1983a). *The eye of the beholder: Measuring aesthetic development*. Unpublished Ed.D. dissertation. Cambridge, MA: Library, Harvard University.
- Housen, A. (1987). Three methods for understanding museum audiences. *Museum Studies Journal*, 2 (4), 41-49.
- Housen, A. (1989). Museum audience research and evaluation. Paper read at Museum Education Division Luncheon, NAEA Conference, Washington, DC, Spring, 1989.
- Housen, A., Miller, N. L., & Yenawine, P. (1991). *Preliminary report, year I (1989-1990)*. New York, NY: MoMA Research and Evaluation Study.
- Kohlberg, L. (1981). *The philosophy of moral development*. San Francisco, CA: Harper and Row, Inc.
- Lasker, H.M. (1976). *Achievement motivation and ego stages: A cross cultural study*. Unpublished Ph.D. dissertation. Chicago, IL: Library, University of Chicago.

Loevinger, J. & Wessler, R. (1976). *Ego development: Conceptions and theories*. San Francisco, CA: Jossey-Bass.

Lowenfeld, V. (1957). *Creative and mental growth*. New York, NY: Macmillan Co.

Parsons, M., Johnston, M. & Durham, R. (1978). Developmental stages in children's aesthetic responses. *Journal of Aesthetic Education*, 12, pp. 83-104.

Parsons, M., Johnston, M. & Durham, R. (1987). *How we understand art: cognitive developmental account of aesthetic experience*. Cambridge, England: Cambridge University Press.