Psychology Research, ISSN 2159-5542
April 2012, Vol. 2, No. 4, 205-221

Degrees of Quality: A Method for Quantifying Aesthetic Impact*

Duane E. Lundy

Indiana University East, Richmond, Indiana, USA

Her sense of taste is such that she will distinguish with her tongue the subtleties a spectrograph would miss.
Brian Eno, *The Fat Lady of Limbourgh*

It is impossible to continue in the practice of contemplating any order of beauty, without being frequently obliged to form comparisons between the several species and degrees of excellence, and estimating their proportion to each other.
David Hume, *Of the Standard of Taste*

An aesthetics rating method facilitating quantitative refinement of individual aesthetic perception was created and applied to a large volume of music works. An exemplar method of concept perception, along with some elements of Thurstone's attitude scale technique, was combined to create the DLAIRM (Definitive Levels of Aesthetic Impact Rating Method). To enhance individual refinement, one must first become familiar with a maximally wide range of works in a particular aesthetic area that are perceived as varying in aesthetic impact from 0 to 100, and then determine one's list of definitive exemplar works for each level of aesthetic impact at gradations of 5%. When evaluating the impact of subsequent works, judgments are made relative to these familiar exemplars along the entire perceived continuum, allowing individual ratings of 2.5% precision. Using this method for a randomly selected group of albums led to more refined results on relevant statistical markers compared to professional critics' ratings of those albums. While it has so far been applied to music, the same process could conceivably be applied to any aesthetic area.

Keywords: rating artworks, aesthetic impact, aesthetic merit or value or quality, aesthetic consensus, critics' judgments, attitude measurement, beauty measurement

Introduction

Aesthetic evaluation has had an interesting and contentious history. Given the existence of such expressions as "Beauty is in the eye of the beholder", and "It is beautiful if someone thinks it is beautiful", it appears that some have thought it pointless to attempt any objective comparison of artworks. These kinds of statements, however, had been applied towards human physical beauty, yet research eventually demonstrated very high interrater consensus in facial beauty ratings (Cunningham, Roberts, Barbee, & Druen, 1995).

* **Acknowledgment**: This research was conducted as an independent project by the author. Some of this material was presented as one portion of an invited talk at the 2010 IGEL biannual conference in Utrecht, Netherlands. The author would like to thank Michael Cunningham for helpful comments about this project.

Duane E. Lundy, School of Humanities and Social Sciences, Indiana University East.

Others have argued that some consensus about artistic beauty should be expected, at least among certain people (Hume, 1757; Kant, 1790; Mothersill, 1984; Reid, 1967; as cited by Manns, 1998). Yet others have argued that only certain aspects can be consensual, such as whether or not the intended artistic purpose was achieved, but that taste is purely subjective (Kaufman, 2002). Despite of the philosophical debate, empirical research has demonstrated that even without agreeing on any particular method of evaluation, reasonable consensus has been established in varied areas of aesthetics, including music (Farnsworth, 1949, 1950; Lundy, 2010a; North & Hargreaves, 1995, 1996), film (Boor, 1990, 1992; Simonton, 2004) and visual art (Child, 1962a; 1962b).

Research thus tends to support the view that at least reasonable consensus in aesthetic evaluation is attainable. But how accurate are these consensus estimates and what exactly is being rated consensually? Aesthetics has been defined generally as: “The study of the feelings, concepts, and judgements arising from our appreciation of the arts or of the wider class of objects considered moving, or beautiful, or sublime” (Blackburn, 1996, p. 8). Aesthetic impact is defined here as a combination of factors that culminates in a greater or lesser overall aesthetic experience for a particular individual when perceiving a relevant stimulus. A serious judge should ideally give a reflective judgment in the context of a wide array of previous works that will be somewhat “disinterested” (Kant, 1790), in the sense of minimizing any personal agendas, non-aesthetic biases, prejudices, etc. (Lundy, 2010b). If each judge were to use a standard method that allowed for high perceptual acuity, then the situation should improve even further. What more could we ever do in an area where there will always be at least some idiosyncratic disagreement?

The philosopher and critic Walter Pater considered “the good life” to include the refinement of one’s aesthetic sensitivity, and believed that in art there is a place for specialized techniques of analysis and criteria of value akin to scientific investigation (Child, 1940). Simply, giving ordinal or interval ratings to artworks is not uncommon (Boor, 1992; DeCurtis, Henke, & George-Warren, 1992; Halsey, 1976; Manaris, Romero, Machado, Krehbiel, Hirzel, Pharr, & Davis, 2005), yet the idea of developing a precise quantitative method to compare works of art seems to have rarely occurred to subsequent scholars. Some scholars have developed tests of aesthetic knowledge and judgment (rather than methods of evaluation), but interestingly, most of the developed tests have been developed for visual art and related perceptual abilities (Child, 1962b; Bulley, 1951; Graves, 1948; Meier, 1942), with no mention of more than one aesthetic domain at a time (i.e., music, film, literature, visual art and so on). Art appreciation itself is supposed to be about enhancement or refinement of one’s aesthetic perception, including aesthetic fluency (the development over time of one’s sensitivity in terms of knowledge base and sophistication in one or more areas of art; Housen, 1992; Parsons, 1989; Silvia, 2007; L. Smith & J. Smith, 2006) and aesthetic judgment (or “good taste”, the extent to which a person’s judgments about aesthetic value ultimately agree with some criterion, such as the consensual judgment of experts; Child, 1964). What is now needed is a parallel leap in rating refinement. The rating method described in this paper

attempts to provide such refinement: a method of aesthetic evaluation that allows for precise quantitative refinement of each individual judge's aesthetic impact perceptions. This would maximize the meaning and precision of a judge's ratings in any of the area of art. In addition, this would ultimately improve the accuracy of aesthetic consensus estimates, an accepted method of determining aesthetic quality (Amabile, 1982). At the very center of this method, there is an answer to a key question in aesthetic evaluation: How does one ever know what a specific numerical rating really means?

Method Development in Relation to Theories of Cognition, Perception and Attitudes

The author has been developing and refining a method of comparing levels of aesthetic impact called the DLAIRM (Definitive Levels of Aesthetic Impact Rating Method). In terms of theory, it fits most closely with exemplar theories of pattern perception and concept formation in cognitive psychology (Medin & Schaffer, 1978; Nosofsky, 1986), along with some elements of Thurstone's (1927) judgment technique of attitude measurement, including his law of comparative judgment. In combination, these turn out to be highly applicable and useful in judging aesthetic works. Theories of human pattern perception are particularly useful to aesthetics, because we are in fact dealing with our aesthetic perception of potentially artistic patterns in various domains: Music represents patterns of notes (and words) recognized via familiarity in repeated listening, literature represents narrative patterns of words, visual art represents patterns of images and films provide narrative patterns in multiple aesthetic modalities (visuals, language, nonverbal behaviors, music, etc.). This is not unlike facial beauty, wherein human beings show predictable reactions based on the immediate perception of facial feature patterns (Olson & Marshuetz, 2005; Cunningham, 1986) and ratings of faces have been shown to be highly consensual across judges (Cunningham et al., 1995). Different artworks can similarly create degrees of aesthetic impact on human perceivers (i.e., differing favorability ratings or hedonic pleasure; Berlyne, 1974; Holbrook & Zirlin, 1985).

For the perception of aesthetic evaluation, as opposed to basic perception, instead of being about how one recognizes what a stimulus is (a concept), it is about how one can recognize the relative aesthetic impact of a stimulus. Exemplar theories of basic perception can thus be applied to aesthetic perception. There are different ways that various aesthetic objects could end up being perceived as equally good or bad. For example, two songs could be of completely different styles but still both perceived as masterpieces. In fact, apart from basic factors such as harmony and complexity (Eysenck, 1957; Manaris et al., 2005), it has been proven very difficult to specify (with any consensus) the necessary and sufficient definitive characteristics that an artistic work must meet to be considered to have a certain level of aesthetic merit (Mothersill, 1984), i.e., that piece X must have these particular Y characteristics to be considered to have Z aesthetic quality. Yet much of the research and discussion in the area of aesthetic evaluation has seemed to focus on just this kind of goal (e.g., see numerous related references listed in Millis & Larson, 2008), and this may be part of the

reason that no evaluation method has ever been widely accepted. A potentially more fruitful focus instead might be on whether people have similar or different experiences of aesthetic stimuli regardless of possible underlying factors. People have shown a lot of consensus in their ratings of facial beauty and have shown to be similarly impacted by it (Hatfield & Sprecher, 1986), despite usually being unaware of exactly what creates the effect. It is the same for aesthetic ratings, we do not necessarily know why people are agreeing, yet we still can know when they are agreeing. In support of this view, without any predetermined definitions or rating methods, a large sample of modern music critic pairs have been shown to be moderately consensual in their independent ratings of albums, with 86% of critic pairs showing positive correlations and not a single pair showing a negative correlation (average $r = + 0.49$; Lundy, 2010a). It would, of course, be nice to know the underlying factors that create consensus, but this does not appear to be essential. Veryzer (1993) found that aesthetic responses to a variety of visual stimuli were consistent with changes in specified design principles across participants (e.g., proportion and unity), but these could not be described at the conscious level, except for vague preferences. He argued that these intuitive preferences must stem from IPAs (internal processing algorithms). The current method capitalizes on such intuitive heuristic aesthetic responses that differ in perceived impact, and is consistent with key characteristics suggested to be inherent in human intuition in general (Seligman & Kahana, 2009).

The proposed method does not require any focus on the difficult (and potentially hopeless) task of getting fellow aesthetes to agree on all of the specific underlying reasons for aesthetic quality. One simply needs to precisely determine one's own intuitive ratings of perceived aesthetic impact using an efficient but precise exemplar-based method; calculating the degree of interrater consensus with other aesthetes would then be more accurate. This method becomes a way of operationalizing a level of aesthetic impact in a particular aesthetic area (i.e., knowing a certain level of aesthetic impact when you "see" it). For instance, on a 0 to 100 scale, what does a 100% song sound like or what does an 80% film look like? This can be accomplished by engaging in a task that passionate aesthetes should find both feasible and enjoyable: coming up with one's own multiple exemplars that are used as typical representations of each level of aesthetic impact (just as two or three photos could be used as exemplars to best represent extremely high facial beauty). Different judges would not even have to use the same particular exemplars in the process of evaluation, as long as independent judges end up with similar numerical ratings of the same artworks. This is akin to Kant's (1790) argument for "objectivity in common subjectivity".

As DLAIRM has been applied to music ratings, there are reliably perceivable levels of aesthetic impact, so that particular songs can serve as exemplars for each numerical level. Each song at each level then serves as a definitive example of that level of aesthetic impact that can be used as a comparison device when attempting to determine the ratings of subsequent songs. The usability of such an exemplar method of definitive songs fits with a statement made by Medin and Schaffer (1978) relating to exemplar theory: "The main idea is that a

probe item acts as a retrieval cue to access information associated with stimuli similar to the probe" (p. 207). The premise is that if this retrieval cueing process is akin to the way our stereotyping brain recognizes concepts, then exemplars arguably provide the easiest, most intuitive, efficient, usable, and potentially accurate way for us to recognize and acknowledge various levels of aesthetic impact. Whatever complex interactions of various artistic dimensions may take place to create an aesthetic experience, the essence of the degree of that experience can be encapsulated by a memorable exemplar—an instance of a certain level of impact that can be reliably recognized.

Some of the ideas in Thurstone's (1927; 1928; 1930) attitude measurement technique become useful for creating a judge's list of aesthetic exemplars. He worked on creating unidimensional scales, wherein one attempts to locate stimuli on single dimensions of favorability (e.g., attitudes towards guns). In the case of aesthetic evaluation, the focus becomes the relative impact of artworks. Thurstone used the idea of scale values, which are given as the most typical psychological reaction to a stimulus. Any difference in judged goodness of one stimulus to another is related to the difference in their scale values on a particular psychological dimension. His method relates to the method of paired comparisons, i.e., judges evaluating which of the two stimuli lies above the other. It also relates to his method of equal-appearing intervals, where judges rank order a group of stimuli along a continuum of equal intervals (Thurstone & Chave, 1929). With aesthetic judgment, there are perceivable levels of aesthetic impact that can be made more concrete and reliable by using exemplars; some exemplars (e.g., songs) are perceived as equivalent in impact and others are perceived to lie above or below others. The key difference from Thurstone's technique is that instead of using a group of judges to determine scale values of attitude statements, an experienced individual judge can precisely determine exemplar aesthetic impact scale values (i.e., actual real-life examples of aesthetic works at each level that are roughly equivalent in perceived impact). Multiple songs (or films, wines, etc.) at the same perceived level may share some characteristic features but they obviously are also different in some features. What they do share is a comparable level of aesthetic impact on a perceiver (e.g., both song A and song B could be examples of level 90 out of 100). Following this kind of exemplar process is arguably the best that we can do to rate aesthetic impact with both maximal reliability and validity. Exemplars are akin to "fuzzy concepts", or categories that cannot be easily defined; they do not have perfectly definable features or boundaries (i.e., the borders are fuzzy). Instead they have characteristic features that all do not have to be possessed in each example of that concept (Rosch, 1975; Smith, 1988), and yet they function very well in everyday perception. In aesthetic works, there are different ways to achieve a certain level of aesthetic impact, and individual raters can determine their own exemplars of each level. Thus, DLAIRM is partly data driven, in the sense that it is created from the ground up, with an individual rater first identifying the full range of exemplars perceived to exist in his or her experience within a certain area of art, and then carefully determining fine-tuned increments in levels of impact (akin to

what Hume called “proportionality”).

This method is also an anchoring technique (Tversky & Kahneman, 1974), in that the chosen exemplars that represent each level of impact serve as anchors (i.e., referents) for future judgments. In other words, a judge’s previous determination of a particular artwork’s level of impact serves as an anchor for subsequent ratings. Precision of the anchor would then influence the amount of adjustment (Janiszewski & Uy, 2008), and arguably, the amount of error or imprecision in the ratings. In attempting to contrast the relative degrees of aesthetic impact among works, the hypothetical ideal for any judge would simultaneously compare every work one has ever come into contact with, but this is obviously not possible in terms of the limits of human memory and conscious attention. Therefore, one needs a technique that preserves the idea of relative contrasting comparisons of aesthetic impact, but that fits the inherent cognitive limitations of the human brain, leaving us with the idea of definitive exemplars. Supporting this idea, Martindale (1988) noted that, “In the perception of music, as in the perception of literature, people seem to extract the gist rather than the details” (p. 23). Similarly, Schwarz (1999) noted that when asked to form an evaluative judgment of a target stimulus, individuals did not retrieve all knowledge that could be relevant, but based their judgment on a subset of information that was most accessible at the time of the judgment, and they “… need a mental representation of a standard against which the target is evaluated” (p. 100). An exemplar method ensures that the central contrasting standard of evaluation is always the same, regardless of what artwork one is rating, thereby always bringing the judge back to the same comparison points for all evaluations, maximizing the chance of stable judgments and minimizing context effects. What become “chronically accessible” are meaningful and proportional pre-ranked exemplars. Consistent with DLAIRM’s memory for exemplars notion, Russell (1986) found that, “… ratings of pleasingness and familiarity obtained under conditions where people are required to remember a recording, cued by artist and title, accord very closely with ratings obtained immediately after the recording has been heard” (pp. 40-41). He called this “cued memory”, and this supports a key claim presented here that intuitively memorable exemplars provide lasting mental benchmarks for contrasting and judging subsequent songs.

How Individual Judges Can Use DLAIRM

In terms of illustrating exactly how actual aesthetic impacts evaluation can be done quantitatively by interested aesthetes, elements of Thurstone’s attitude measurement technique, in combination with an exemplar theory of perception, are now summarized and applied specifically to music appraisal.

Compile and Rank-Order a Maximally Wide-Ranging Pool of Artworks

For one’s aesthetic area of interest, compile an extensive pool of items (e.g., songs or films) varying in degree of aesthetic impact (what Thurstone called “favorability”) along a single continuum from 0 to 100, which is large enough to represent as many different points along the scale as possible, including neutrality (i.e., 50), maximal badness (i.e., 0 or maximum negative valence), and maximal goodness (i.e., 100, or maximum

positive valence or “profoundness”, a term that others have also used to mean the highest level of music quality, e.g., Holbrook & Zirlin, 1985; Levinson, 1992; White, 1992). Manaris et al. (2005, p. 64) similarly used 50 to represent aesthetic neutrality (“emotional indifference or neutral reaction”).

One needs to compare and contrast artworks, deciding which one lies above or below each other in perceived aesthetic impact. Thus, ideally the entire possible range of scale values is represented (and can be thought of as easily understood “percentages”), where for each individual judge, the lowest ratings would represent the worst work(s) he/she has ever experienced and the highest ratings would represent the best work(s) ever experienced. Note that this only works well when an extremely wide range of works is experienced, and this relates to a key feature of the face validity of any scale, wherein the full universe of behaviors is represented (Aiken & Groth-Marnat, 2006). Holbrook and Zirlin (1985) came closest to identifying this range of possible aesthetic responses, but still mentioned less than half of the rating scale (pleasant to profound), in effect omitting neutrality and the bottom half of the scale (i.e., degree of badness/unpleasant aesthetic impact/negative valence). Ignoring the opposite of beauty has apparently been a consistent oversight into aesthetics; Lorand (1994) had asserted that there are at least six different ways that art can be bad. Any of these could be incorporated into a sub-50 aesthetic impact numerical rating. When a judge gives a rating below 50 in DLAIRM, the implication is that there are more bad aesthetic elements in the artwork than good elements.

In applying this method specifically to modern music evaluation, thousands of songs were gathered from hundreds of musicians varying in perceived aesthetic impact in multiple pop/rock subgenres over a 40-year time span (1960 to 1999). This included an ongoing search for the best and worst possible songs that could be found, as well as appropriate examples of songs perceived to be fairly neutral (i.e., neither good nor bad, or having good and bad elements in roughly equal amounts). Thus, this method was developed in an area in which the author has the greatest interest and knowledge (pop/rock music), but the same process should be applicable to any domain or subdomain of aesthetics. For someone else the focus could be classical music, films, paintings or wine-tasting, etc..

Identify “Definitive Works”: Memorable Exemplars at Equal Intervals Along the Continuum

As one attempts to put the various artworks in rank order, use particularly memorable ones as definitive exemplars of particular levels. The author attempted to put various songs in order of perceived aesthetic impact from neutral up to maximum positive impact (“100”) and from neutral down to maximum negative impact (“0”) in equal intervals of 5% (or 21 gradations) out of 100 (i.e., 0, 5, 10... 50… , 90, 95, 100), effectively creating a list referred to as “definitive songs” (i.e., definitive exemplars or levels). The author did this with one judge (the author himself), but this conceivably could be done by any experienced judge independently doing the same thing. The goal is to find the best overall exemplars (i.e., simplest, most memorable works). In this way, reasonably high precision within an individual rater’s aesthetic impact

ratings can be achieved. At least two exemplars need to be used for each level to ensure that it is not the specific characteristics of one particular artwork that is being represented, but instead their shared level of perceived aesthetic impact (e.g., for pop/rock music, the author ended up using John Lennon's *God* and Brian Eno's *Golden Hours* to represent level 100). These songs are top tier exemplars that the author personally used, and whether others would agree would be an empirical question, although both of these songs do come from consistently highly rated albums by professional critics (Lundy, 2010a). Moreover, across the songs that were used as exemplars, the average correlation between ratings and the ratings of Rolling Stone critics' ratings from the corresponding albums was + 0.87 (DeCurtis, Henke, & George-Warren, 1992; Marsh & Swenson, 1979, 1983).

Use Your "Definitive Exemplars" List to Rate All Subsequent Works With High Perceptual Acuity

One then uses a definitive songs list to rate subsequent songs by comparing each with the exemplars, and gives each song a rating at one of these levels, or in the case of indecision, in between the two levels. This last technique fits with a general grading recommendation made by Cronbach, Bradburn, and Horvitz (1994), who suggested that raters be encouraged to use intermediate scale values for students' borderline responses. This yields 41 possible ratings for any song (i.e., 2.5% gradations). To rate albums as a whole, one simply has to add up each song rating on the album and divide by the number of songs to get the average level of aesthetic impact for that particular album, which could vary from 0 to 100 (e.g., the author's compiled pop/rock album impact ratings currently vary from a low of 9% to a high of 93%). Once everyone "purifies" one's own determinations, i.e., sharpens one's skills at precisely quantifying aesthetic evaluations at high reliability and maximum discriminability, measuring interrater consensus then becomes maximally meaningful and precise.

Potential Problems

The possibility arises that an evaluation system with a high number of rating gradations could end up being arduously difficult to use effectively and yield low test-retest reliability. Myford (2002) compared rating scales of various gradations for evaluating student artwork and concluded that reliability did not increase with more than five rating gradations. Although that rating process did involve utilizing visually displayed exemplars along a continuum of quality (for visual art), these were not created by each individual judge who had been practicing perceptual acuity over a long period of time, but were defined by a rating trainer, both in terms of their relative position on the continuum and what features one should look for. Thus, although the raters had some general arts background, each rater had not personally invested a lot of time to develop high familiarity and practice perceptual acuity for himself or herself (as would be the case with DLAIRM). Moreover, the rating scales used were graphic scales with a visual continuum where one indicated one's rating with a slash, whereas precise numbers were not used on the scales (compared to 0 to 100 at 2.5% gradations in

DLAIRM). It is not surprising then that higher reliability was not obtained with such scales of 5 *vs.* 10 gradations. Other music rating studies had already successfully used 13-point (Heyduk, 1975) and 14-point (Krugman, 1943) rating scales. Moreover, in the author's study of modern music critics, moderate consensus was obtained while many of the critics used 11-point scales (Lundy, 2010a). Payne (1997) suggested that depending on the sophistication level of the raters, scales having as many as 20 categories could be used (DLAIRM starts with 21). Overall, evidence suggests that it can be tough to use numerous gradations at any higher reliability than fewer gradations. However, the focus should ultimately be on validity, and it is tough to argue against the notion that numerous gradations of aesthetic quality in artworks do in fact exist. Thus, as long as each individual rater's test-retest reliability does not decrease substantially as precision increases, then more is better in terms of refining aesthetic perceptual acuity. However, without a method like DLAIRM, more than 11-13 levels of precision could be expected to be very difficult to use reliably, but that is a key point. What becomes important is not so much what a typical person easily does do, but what a trained and knowledgeable person can do becomes maximally precise and accurate. When anyone who is serious about accuracy has sifted through thousands of artworks, and looks at definitive exemplars closely, it becomes very obvious that a relatively small number of rating levels are simply not precise enough to accurately reflect reliably perceivable variation in the impact of aesthetic works. In keeping with the Thurstone's model, the bottom line is that if one can reliably place one work above or below another than there is a perceivable difference in aesthetic impact, and the rating scale being used needs to be able to document such differences so that all ratings are maximally, proportionally meaningful (thereby bringing an aesthetic judge closest to Hume's "proportionality" goal). What one needs is a method of maintaining reliability as the number of gradations increases as much as possible, and this is where identifying one's own definitive exemplars along the full continuum of a given aesthetic domain is critical. In other words, using the definitive levels technique gives us a way of potentially reaching very high individual rating precision.

Note that with the use of a precise rather than rough rating scheme, when one makes "errors in judgment" about the level of aesthetic impact of an artwork, they are going to tend to be smaller. For example, one's ratings could be three levels of another judges' ratings on DLAIRM (i.e., a 7.5% difference), and still be almost twice as precise as two judges who were only off by one level on a seven-point rating scale (i.e., a 14% difference). With DLAIRM, two albums that differ by 2.5% (e.g., 62.5% *vs.* 60.0%) are actually quite different; this difference implies that, on average, every single song on the first album is one level higher than every single song on the second album. Data collected on the author's use of this method show that more than 5% individual precision can be reliably achieved.

Preliminary DLAIRM Data

Data compiled so far that clearly document the usefulness of this method are biographical in nature. In

an Ebbinghaus self-study fashion, the author has studied his own use of the method in great detail, applying the method to a large volume of rock/pop music works released over a 40-year span (between 1960 and 1999). This approach is especially well-suited to testing DLAIRM, because the whole point of this method is not generalizing in the usual sense, but to maximize the rating precision within each individual. The use of individual ratings is also not unprecedented in music evaluation. Halsey (1976) provided his own ratings of classical music that he correlated with other measures. Relevant and supportive data so far have come from five sources. First, to demonstrate general soundness in the author's overall music evaluation tendencies while using DLAIRM (i.e., aesthetic judgment), he compared his ratings of a random sample of 364 albums with a database of professional critic ratings that he had compiled for another study (Lundy, 2010a). Album ratings showed significant positive correlations with six out of nine critics with no negative correlations (average $r = + 0.33$, range = + 0.11 to + 0.63). Second, as mentioned earlier, the definitive exemplars that the author used correlated very highly with published ratings given by Rolling Stone critics for the albums that the songs came from ($r = + 0.87$, $p < 0.0005$, $n = 29$). Third, the change in the author's blind re-ratings across 333 randomly selected albums released between 1960 and 1999 was measured. Demonstrating extremely high test-retest reliability, the average change in the album ratings over time was extremely small at 1.06%, even though the time lag between the two ratings tended to be extremely long, ranging from four to seven years. Fourth, another measure of test-retest reliability was done with a random sample of 395 songs (one album chosen randomly each year from 1965 to 1999). With an average time lag of 20 months, these songs were rerated blindly using only a definitive songs (exemplars) list, and it was found that test-retest reliability was once again extremely high ($r = + 0.93$, $p < 0.0005$), and individual song ratings only changed an average of 2.68% (with 92% of the rating differences falling between 0% and 5%). Fifth, statistical markers of quantitative refinement identified in a recent study among professional music critics were analyzed (Lundy, in press). This study found support for a prediction made that ratings of a large number of modern music albums by professional critics would follow a roughly normal distribution. Some critics' rating distributions were especially close to normality. An assumption of DLAIRM is that proper proportionality of one's ratings can only be fully realized by identifying and contrasting artwork exemplars (e.g., songs) across the entire range of perceived aesthetic quality (i.e., from worst experienced to best experienced exemplars). This process was expected to produce a highly normal distribution of ratings. In fact, the use of DLAIRM for the author's ratings of a random selection of albums ($N = 364$) fit the normality expectation more closely than for any of the professional critics, with near zero skewness and kurtosis (even though the author's main goal has been to find the greatest works, which is why the mean is higher than 50). The rating distribution was not only mound shaped, but this occurred at a much higher level of precision, and unlike all of the 31 professional critics studied, it was not significantly different from a perfectly normal distribution (Kolmogorov-Smirnov (K-S) = 0.03, $p = 0.20$; see Figure 1).

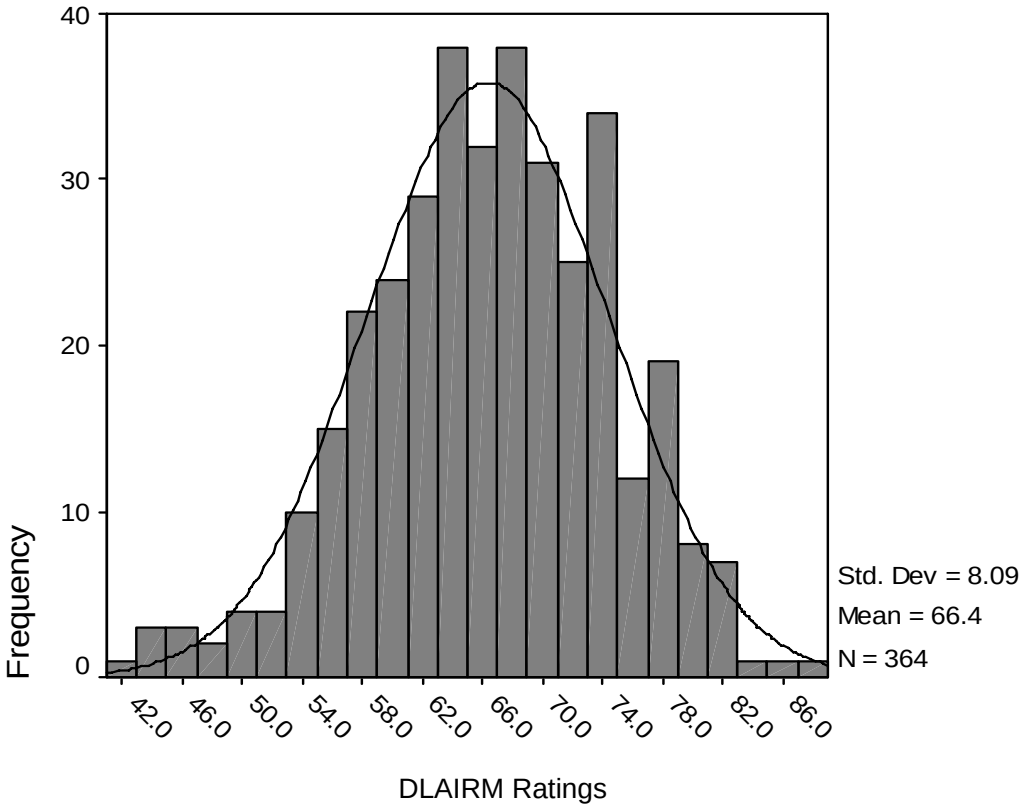

Figure 1. Frequency distribution of album ratings using a precise, exemplar rating method (DLAIRM).

In sum, these forms of data all support the notion that DLAIRM can be used one experienced judge at a time to sharpen one's skill at precisely quantifying one's ratings, which has been demonstrated effectively for "Judge 1". The 2.7% average song rating change clearly shows that a scale with fine-grained gradations can be reliably used, i.e., well within the 5% exemplar gradations (and close to the intended 2.5% precision) in the definitive songs list used to determine the ratings. If a judge was only able to achieve blind test-retest reliability at 9%, then no more than an 11-point scale could be used (or if 8%, then only a 13-point scale could be used, etc.). Yet the reliability data, for music at least, were much more precise than this. In contrast, professional music and movie critics tend to use scales with only 10% to 20% precision.

The kinds of biographical data analyses employed here could be used in general by experienced individual judges to "self-test" both their use of this method and their aesthetic judgment tendencies. These types of data are not typical in psychological research but can be useful for one judge at a time, and the goal of DLAIRM is to do just this—to make an individual reliably precise in his/her aesthetic impact ratings. At least two things relating to this method remain to be seen: (1) Whether other individual judges can achieve similarly high levels of precision for music ratings (and to what extent a high level of preexisting aesthetic fluency is required to do this); and (2) Whether the 5% gradations used to create the definitive level

exemplars for music is achievable for individual judges in other aesthetic domains (e.g., films, visual art, or literature).

Key Advantages of DLAIRM

The following strengths of DLAIRM include: (1) Sensory and perceptual contrast information is provided over the full range of a particular aesthetic modality (i.e., songs varying from roughly 0 to 100 in aesthetic impact), thereby giving one maximal aesthetic perspective; (2) It takes advantage of our intuitive perceptual pattern abilities; (3) It allows for measurable test-retest reliability when a judge blindly rerates the same works; (4) Numerous rating gradations allows for high precision of perceptual acuity within an individual judge's ratings; and (5) It makes efficient use of familiar memory cues (e.g., well-known song exemplars), provided at varying aesthetic impact levels. Overall, when using DLAIRM, an operationalization of each level of aesthetic impact is always provided (e.g., knowing each level of music aesthetic impact when you hear it). In contrast, the following weaknesses exist in critical quantitative judgments of music and film (Lundy, 2010a): (1) low precision (10% to 20% rating gradations), (2) ceiling effects among some critics (i.e., overrating, mistakenly implying that greatness is commonplace), and (3) floor effects among a few critics (i.e., too many "bomb" or "zero" ratings, which should be extremely rare). Such imprecise and non-proportional judgments do a disservice to truly great artists as well as would be aesthetes searching for aesthetic guidance. Following the DLAIRM process could help to minimize such problems among professional critics, as well as a longer list of potential problems in non-professionals, such as lack of exposure to a diverse range of aesthetic stimuli. If a particular judge had never been exposed to works near the highest level of aesthetic merit (i.e., 95-100), he/she would likely perceive the "top" to simply be the highest level he/she had heard so far, which might be as an low as an expert's 70 or 75. Consistent with this, laypersons have been found to give more top tier ratings compared to experts (Boor, 1992; Baer, Kaufman, & Gentile, 2004; Holbrook & Zirlin, 1985). Moreover, the importance of judges being exposed to the full range of quality in an aesthetic domain was mentioned in a study comparing experts and non-experts who were judging poems: "… experts probably gave lower average ratings of the poems (than non-experts), because they generally read much higher quality poems than the undergraduate ones used" (Kaufman, Baer, Cole, & Sexton, 2008, p. 175). Another advantage of DLAIRM is that it helps to minimize a potential problem with the Thurstone's attitude continuum method (Edwards, 1957): How does one know if the space between the rating intervals is equal? The key is to make the gradations as fine-grained as possible, yet still making such distinctions with reasonable test-retest reliability. In addition, DLAIRM also addresses a specific problem observed in most modern music critics' rating distributions, which is too many ratings between 70 and 90 (Lundy, in press); DLAIRM forces one to enhance one's perceptual acuity, making it maximally quantitatively proportional, so that one can distinguish more clearly among modest achievements, great

achievements and masterpieces.

The method described here is completely consistent and complementary with Amabile's (1982) CAT (consensual assessment technique) for judging creativity, wherein multiple judges rate the same works and consensus rates are determined. Previous research using the CAT has established good interrater reliability among judges of both creativity and aesthetic quality (e.g., Baer et al., 2004), but a problem is that no standard and precise quantitative rating method has ever been employed by all judges. This makes determinations of objective aesthetic quality imprecise and lacking in meaning. For example, how does a judge know what his/her poem, song or movie rating of "85" really means, unless it is determined by comparing and contrasting the quality across the full range of existing poems, songs or movies? If each judge could first maximally refine his/her own estimate of aesthetic impact by using DLAIRM, then the group estimate of aesthetic impact would become maximally accurate and meaningful. In short, Amabile's method for creativity judgment specifies the importance of consensus and how it can be calculated, but no standard, refined rating method is used. Thus, DLAIRM is not in competition with CAT; DLAIRM's goal is to make individual ratings of creativity or aesthetic impact maximally precise and valid, before comparing them with others' ratings (via CAT). Such a process would also enhance the meaningfulness of exact agreement measures of judges' consensus, such as the extent to which judges give exactly the same ratings to the same songs. DLAIRM should be viewed as a "pre-CAT necessity" for each judge.

Future Directions

Overall, DLAIRM is a way of standardizing aesthetic impact ratings, achieved by providing in essence a frame of reference for comparison to preexisting ratings of all other artworks rated (i.e., one has prior *norms* based on a large sample, or a normative group with which to compare). It is only in relative contrast to previously experienced aesthetic stimuli covering a wide range that quantitative estimates can be made with precision and proportionality. Note that in general, "raw scores" are not precisely meaningful unless they are compared to overall norms, as is true for personality measures (Costa & McCrae, 1992). In the aesthetic arena, these norms would simply be all previously rated songs, films, or novels, etc.. When one has data driven definitive artworks, then and only then, do the numbers used for ratings gain some real meaning? They become operationalized and relative. Then one gains some feeling for what a rating of 60, 80, or 90 really means. Then, just as a person's standing on an IQ test can be evaluated by comparing his performance with a norms table, the same could be done to evaluate a film's estimated relative standing compared to the distribution of past film ratings among refined judges. Ratings could also be transformed to standard *z*-scores and percentiles. One could also most accurately measure the frequency distribution of the aesthetic impact of released aesthetic works (e.g., there is some evidence for a roughly normal distribution; Lundy, in press). Another application would then be that accolades and awards based on expert ratings, as well as "best of all

time" determinations, would become maximally valid, making them more consistently well-deserved and meaningful.

But not everyone's opinion can be counted; there must be prerequisites of motivation, experience and refinement (Child, 1962a, 1962b; Eysenck, 1940, 1941, 1957, 1972; Hume, 1757; Lundy, Schenkel, Akrie, & Walker, 2010; L. Smith & J. Smith, 2006). This would include background experience such as wide ranging familiarity, but also rating behavior incorporating the use of a precise method of evaluation. It would be interesting to have some critics try this method and see how it works for them, compared to whatever it was they were doing before, and at the same time, it would also be useful to find out the specifics about what they are actually doing. Kaufman et al. (2008) suggested that expert judges might still use inappropriate standards or have a biased agenda. The general usefulness and validity of DLAIRM could also be tested by finding out just how much background familiarity and sophistication (aesthetic fluency and judgment) is necessary to make rough versus precise reliable and consensual distinctions. The bottom line, however, is that a person could be very high in essential characteristics of good critics, such as aesthetic fluency, intelligence, etc., and would still not rate the songs with sufficient precision and reliability without a method like DLAIRM.

The subjective, more personally idiosyncratic, non-consensual elements of aesthetics may well be shown to comprise a relatively small portion of aesthetic impact evaluation compared to the consensual elements, when evaluated by refined independent judges using a standard and precise rating method, such as DLAIRM. Such a method is essential to most accurately test this prediction, and it is suggested that all serious critics need to make sure they are being consistent and proportional within their own aesthetic viewpoint. The author's own song ratings, for instance, changed as much as 25% from initial ratings once a full range of exemplars was utilized. Much heated and longstanding disagreement and debate about artworks may eventually be seen to be about small degrees of difference, because of definitive gradations and an overall perspective of the range of said gradations, people are free to endlessly argue about relatively slight degrees of difference. This is to say that sophisticates would probably agree about groups of artworks that were at or near the top tier in a given artistic area. For example, even if some of us disagree about which is the better album, U2's *The Joshua Tree* or Jimi Hendrix' *Are You Experienced*? We are probably just splitting hairs at this point. Related to this idea, Rosen (1981) suggested that at the top tier in any field, small differences in quality tend to be artificially magnified. One thing that would be interesting to investigate is whether some aesthetic domains, such as music, are more or less consensual than other domains (e.g., film or visual art). By utilizing DLAIRM across all domains, one could obtain such meaningful comparisons. Interested readers could use this method for themselves in their aesthetic areas of interest, attempting to determine their definitive level exemplars of aesthetic impact, and then share their results. The time has come for individual aesthetes to quantitatively refine their aesthetic perceptual acuity in their own personal universes.

References

Aiken, L. R., & Groth-Marnat, G. (2006). *Psychological testing and assessment* (12th ed.). Needham Heights: Allyn & Bacon, Inc..

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology, 43*, 997-1013.

Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal, 16*, 113-117.

Berlyne, D. E. (1974). *Studies in the new experimental aesthetics*. New York: Wiley.

Blackburn, S. (1996). *Oxford Dictionary of Philosophy*. New York: Oxford University Press.

Boor, M. (1990). Reliability of ratings of movies by professional movie critics. *Psychological Reports, 67*, 243-257.

Boor, M. (1992). Relationships among ratings of motion pictures by viewers and six professional movie critics. *Psychological Reports, 70*, 1011-1021.

Bulley, M. H. (1951). *Art and everyman* (Vol. I). London: Batsford.

Child, I. L. (1962a). Personal preferences as an expression of aesthetic sensitivity. *Journal of Personality, 30,* 496-512.

Child, I. L. (1962b). *A study of esthetic values*. New Haven, C. T.: Yale University.

Child, I. L. (1964). *Development of sensitivity to esthetic values.* New Haven, C. T.: Yale University.

Child, R. C. (1940). *The aesthetic of Walter Pater*. New York: MacMillan.

Costa, P. T., & McCrae, R. R. (1992). *NEO PI-R professional manual*. Lutz, F. L.: Psychological Assessment Resources, Inc..

Cronbach, L. J., Bradburn, N. M., & Horvitz, D. G. (1994, July). *Sampling and statistical procedures used in the California Learning Assessment System*. Report of the Select Committee. Palo Alto, C. A.: Author.

Cunningham, M. R. (1986). Measuring the physical in physical attractiveness: Quasi-experiments on the sociobiology of female facial beauty. *Journal of Personality and Social Psychology, 50,* 925-935.

Cunningham, M. R., Roberts, R., Barbee, A. P., & Druen, P. B. (1995). "Their ideas of beauty are, on the whole, the same as ours": Consistency and variability in the cross-cultural perception of female physical attractiveness. *Journal of Personality and Social Psychology, 68*, 261-279.

DeCurtis, A., Henke, J., & George-Warren, H. (Eds.) (1992). *Rolling stone album guide* (3rd ed.). New York: Random House.

Edwards, A. L. (1957). *Techniques of attitude scale construction*. New York: Appleton-Century-Crofts.

Eysenck, H. J. (1940/1941). The general factor in aesthetic judgments. *British Journal of Psychology, 31*, 94-102.

Eysenck, H. J. (1957). *Sense and nonsense in psychology*. Baltimore, M. D.: Penguin.

Eysenck, H. J. (1972). Personal preference, aesthetic sensitivity and personality in trained and untrained subjects. *Journal of Personality, 40*, 544-557.

Farnsworth, P. R. (1949). Agreement with the judgments of musicologists as a measure of musical taste. *Journal of Psychology, 28*, 421-425.

Farnsworth, P. R. (1950). *Musical taste: Its measurement and cultural nature*. Stanford: Stanford University Press.

Graves, M. (1948). *Design judgment test*. New York: Psychological Corporation.

Halsey, R. S. (1976). *Classical music recordings for home and library*. American Library Association.

Hatfield, E., & Sprecher, S. (1986). *Mirror, mirror: The importance of looks in everyday life*. Albany: SUNY Press.

Holbrook, M. B., & Zirlin, R. B. (1985). Artistic creation, artworks, and aesthetic appreciation: Some philosophical contributions to nonprofit marketing. *Advances in Nonprofit Marketing, 1*, 1-54.

Housen, A. (1992). Validating a measure of aesthetic development for museums and schools. *ILVS Review, 2*, 213-237.

Hume, D. (1757/1965). Of the standard of taste. In *On the standard of taste and other essays*. Indianapolis: The Bobbs-Merrill Company.

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science, 19*, 121-127.

Kant, I. (1790/1949). The critique of judgment. In C. J. Friedrich (Ed.), *The philosophy of Kant: Immanuel Kant's moral and political writings*. New York: The Modern Library.

Kaufman, D. A. (2002). Normative criticism and the objective value of artworks. *Journal of Aesthetics and Art Criticism, 60*, 151-166.

Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal, 20*, 171-178.

Levinson, J. (1992). Musical profundity misplaced. *Journal of Aesthetics and Art Criticism, 50*, 58-60.

Lorand, R. (1994). Beauty and its opposites. *Journal of Aesthetics and Art Criticism, 52*, 399-406.

Lundy, D. E. (2010a). A test of consensus in aesthetic evaluation among professional modern music critics. *Empirical Studies of the Arts, 28*, 243-258.

Lundy, D. E. (2010b). *Avoiding nonaesthetic biases in aesthetic appraisal*. Manuscript in preparation.

Lundy, D. E. (in press). Critiquing the critics: Statistical analysis of music critics' rating distributions as a measure of individual refinement. *Empirical Studies of the Arts*.

Lundy, D. E., Schenkel, M. B., Akrie, T. N., & Walker, A. M. (2010). How important is beauty to you? The development of the Desire for Aesthetics Scale. *Empirical Studies of the Arts, 28*, 73-92.

Manaris, B., Romero, J., Machado, P., Krehbiel, D., Hirzel, T., Pharr, W., & Davis, R. B. (2005). Zipf's law, music classification, and aesthetics. *Computer Music Journal, 29*, 55-69.

Manns, J. W. (1998). *Aesthetics*. Armonk, N. Y.: M. E. Sharpe, Inc..

Marsh, D., & Swenson, J. (1979). *The rolling stone record guide*. New York: Random House.

Marsh, D., & Swenson, J. (1983). *The new rolling stone record guide*. New York: Random House.

Martindale, C. (1988). Aesthetics, psychobiology, and cognition. In F. H. Farley, & R. W. Neperud (Eds.), *The foundations of aesthetics, art, and art education* (pp. 117-160). New York: Praeger.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Meier, N. C. (1942). *The meier art tests. I. Art judgment; Examiner's manual*. Iowa City, I. A.: Bureau of Education Research, University of Iowa.

Millis, K., & Larson, M. (2008). Applying the construction-integration framework to aesthetic responses to representational artworks. *Discourse Processes, 45*, 263-287.

Mothersill, M. (1984). *Beauty restored.* Oxford: Clarendon.

Myford, C. M. (2002). Investigating design features of descriptive graphic rating scales. *Applied Measurement in Education, 15*, 187-215.

North, A. C., & Hargreaves, D. J. (1995). Eminence in pop music. *Popular Music and Society, 19*, 41-66.

North, A. C., & Hargreaves, D. J. (1996). Affective and evaluative responses to the arts. *Empirical Studies of the Arts, 14*, 207-222.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57.

Olson, I. R., & Marshuetz, C. (2005). Facial attractiveness is appraised at a glance. *Emotion, 5*, 498-502.

Parsons, M. J. (1989). *How we understand art: A cognitive developmental account of aesthetic experience*. Cambridge: Cambridge University Press.

Payne, D. A. (1997). *Applied educational assessment.* Belmont, C. A.: Wadsworth.

Rosch, E. H. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General, 104*, 192-233.

Rosen, S. (1981). The economics of superstars. *American Economic Review, 71*, 845-858.

Russell, P. A. (1986). Experimental aesthetics of popular music recordings: Pleasingness, familiarity and chart performance. *Psychology of Music, 14*, 33-43.

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*, 93-105.

Seligman, M. E. P., & Kahana, M. (2009). Unpacking intuition: A conjecture. *Perspectives on Psychological Science, 4*, 399-402.

Silvia, P. J. (2007). Knowledge-based assessment of expertise in the arts: Exploring aesthetic fluency. *Psychology of Aesthetics, Creativity, and the Arts, 1*, 247-249.

Simonton, D. K. (2004). Film awards as indicators of cinematic creativity and achievement: A quantitative comparison of the Oscars and six alternatives. *Creativity Research Journal, 16*, 163-172.

Smith, E. E. (1988). Concepts and thought. In R. J. Sternberg & E. E. Smith (Eds.), *The psychology of human thought* (pp. 19-49). New York: Cambridge University Press.

Smith, L., & Smith, J. (2006). The nature and growth of aesthetic fluency. In P. Locher, C. Martindale, & L. Dorfman (Eds.), *New directions in aesthetics, creativity and the arts* (pp. 47-58). Amityville, N. Y.: Baywood.

Sternberg, R. J. (2006). *Cognitive psychology* (4th ed.). Belmont, C. A.: Wadsworth.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*, 273-286.

Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529-554.

Thurstone, L. L. (1930). A scale for measuring attitude toward the movies. *Journal of Educational Research, 22*, 89-94.

Thurstone, L. L., & Chave, E. J. (1929). *The measurement of attitude*. Chicago: University of Chicago Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Veryzer, R. W. Jr. (1993). Aesthetic response and the influence of design principles on product preferences. *Advances in Consumer Research, 20*, 224-228.

White, D. A. (1992). Toward a theory of profundity in music. *Journal of Aesthetics and Art Criticism, 50*, 23-34.