2009

The Effect of Performance Errors on Perceptions of Performance Quality in J.S. Bach's Bourée from the Suite for Unaccompanied Cello #3

R. Eric Simpson

FLORIDA STATE UNIVERSITY

COLLEGE OF MUSIC

THE EFFECT OF PERFORMANCE ERRORS ON PERCEPTIONS OF PERFORMANCE QUALITY IN J.S. BACH'S *BOURÉE* FROM THE *SUITE FOR UNACCOMPANIED CELLO #3*

By

R. ERIC SIMPSON

A Dissertation submitted to the
College of Music
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Degree Awarded:
Summer Semester, 2009

The members of the committee approve the dissertation of R. Eric Simpson defended on June 15, 2009.

_____
Clifford K. Madsen
Professor Directing Dissertation

_____
Alexander Jiménez
Outside Committee Member

_____
Steven Kelly
Committee Member

_____
Patrick Dunnigan
Committee Member

Approved:

_____
Don Gibson, Dean, College of Music

The Graduate School has verified and approved the above-named committee members.

This is dedicated to Allison Margaret Simpson, my daughter,
who was there while I wrote every word,
reminding me of what really mattered.

ACKNOWLEDGEMENTS

I wish to thank Brady C. Beard, Audio Engineer-Nashville, TN, for his assistance in preparing and editing the audio excerpts used in the study. I further wish to extend my gratitude to fellow graduate student Sharon Graham for her assistance during the data collection portion of this study.

I also wish to express my appreciation to the tremendous colleagues with whom it has been my pleasure to pursue graduate study. I was fortunate to work with a talented group of graduate music education students at Florida State University and the University of Cincinnati College-Conservatory of Music, and their contributions to this piece of scholarship through conversation, interaction, and inspiration are immeasurable.

Thanks to Clifford Madsen, who served as my major professor and directed this dissertation. It is rare to find a truly brilliant mind who is also a genuinely nice person, but I was lucky enough to have both in Professor Madsen. The *joie de vivre* that he brings to his work everyday is truly inspiring, and if I turn out to be half the scholar he is I will consider myself very blessed.

Finally, this manuscript, this degree, and this experience would have never happened had it not been for the continuous support and encouragement of my wife, Meredith B. Simpson. She has provided an anchor for me emotionally, professionally, and even financially throughout my graduate studies. There is no way to adequately thank her for what she's done and continues to do for me, but suffice it to say that every day I realize again how lucky I am to be married to such a smart, strong, kind, beautiful person.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The purpose of this study was to determine how errors in performance affect perceptions of performance quality by musicians while listening to an excerpt of Bach's *Bourrée* from the *Suite for Unaccompanied Cello #3*. Subjects (N=129) heard five excerpts that had been digitally altered to make it appear that the performer was playing wrong notes. Subjects heard the excerpts in one of four randomized orders. Subjects were instructed to rate the quality of the performance by manipulating the Continuous Response Digital Interface. Subjects were also asked to make comments about the listening task in a free-response posttest questionnaire. Mean ratings of each of the excerpts were charted graphically, both as individual excerpts and within each order. Results from the study indicated that negative stimuli elicited responses of much greater strength than did positive stimuli. These responses were much more abrupt than responses to positive stimuli, which tended to be more gradual. Results also indicated that recovery from negative stimuli was slow, often persisting long after the stimulus had passed.

CHAPTER 1
INTRODUCTION

Making good choices is a skill of considerable value. From a very young age, much time, energy, and effort goes into teaching children to discriminate effectively. In many ways, the entire process of education is simply a process of learning to discriminate. Making proper choices is the key to success in many developmental aspects of life; whether choosing the correct block for the hole, deciding which integers will complete the equation, or choosing the correct mate. Many of the most prominent thinkers in the annals of human history have dedicated their lives to decision-making. The philosophy of Plato is concerned primarily with the ideal, of choosing the Form of the Good over the shadows which comprise physical objects. A central theme of the Christian faith is man's possession of free will, and the inherent pitfalls that come with choice; while Jean-Paul Sartre said, "it is only in our decisions that we are important."

Given the seminal position of discrimination in human existence, it is not surprising that decision-making has been a major scholarly interest for some time. In some ways, the entire field of psychology could be said to be obsessed with discrimination. Freud's early attempts at penetrating the subconscious were intimations at shedding light on why people make the choices they make, and each new approach to psychology has followed suit to some degree. Thus, early behaviorists argued that choices are informed by responses to stimuli, while cognitive psychologists believed that choices could be attributable to internal mental processes (belief, motivation, etc.).

Trying to understand why people make the choices they make is a major epistemological puzzle. And, like most good questions, it has led to more questions than answers. Do we choose based on logic? How susceptible to influence are our decisions? Can we really choose, or are our choices already predetermined by our circumstances? Are we, at any time, anything more than the sum of our decisions?

Recent research has begun to suggest a new paradigm for understanding how people make decisions. Advances in neuroscience have presented a clearer picture of how quickly the brain works. Early models of cognitive function, which tended to view the brain as a recording/recall device, are being reconsidered as new scientific advances shed light on how the brain processes information. Because motor information, sensory

information, and cognitive information are processed by 200 billion neurons to contribute to the process of discrimination, much discrimination occurs on a level that does not require cognitive thought. To walk, for example, it is not necessary to make the conscious decision to life one's foot-that task is handled by motor neurons, just as proprioception maintains an awareness of the location of one's feet independently of cognitive functions. Such advances have led researchers to question the nature of the decision-making process. Conventional wisdom holds that people make decisions by evaluating a range of options and choosing the best alternative. Other researchers have demonstrated, however, that discrimination can be affected by a range of factors (prejudice, visual cues, etc.) that may influence decision-making underneath the conscious level.

For the musician, few skills are as important as the ability to discriminate. Musicians are constantly in the process of discrimination, whether choosing the right note to complete a phrase, choosing which composition to perform, or choosing which performance to emulate. Not surprisingly, music researchers have begun to demonstrate that the same knowledge that is redefining traditional decision-making is at play in the field of music. Recent research in preference and adjudication has suggested that musicians' decisions can be influenced by a variety of musical (tempo, loudness, timbral, etc.) and non-musical factors (race, gender, attractiveness, etc.) While some of these factors may be cognitively processed, others could be said to be "flying under the radar."

One of the major obstacles to understanding the discrimination process of musicians has been time. For many years, the only way to gather information about musicians' decisions was after a musical event had occurred. These post-hoc evaluations, known as static measures, could not always be trusted from a validity perspective. Because music occurs temporally, static measures often constituted a "snapshot" of what was essentially a fluid event. Innovations in computer technology, however, have reconfigured the data collection process in the field of music. Continuous measures, which take samples across time, have become increasingly more widespread as microcomputers have become more prevalent and accessible.

One example of this is the Continuous Response Digital Interface (CRDI), a device pioneered by the Center for Music Research, which consists of a potentiometer connected to a microcomputer. The CRDI allows for time-sampling at minute intervals,

allowing researchers to examine how musicians discriminate across time, rather than at discrete moments. This device has many applications, and has been used to examine subjects as diverse as preference, evaluation, teaching, various performance aspects, etc. One of the major thrusts of CRDI research has been in the area of aesthetic experience. Researchers examining aesthetic experience from a continuous perspective have shed new light on how people respond to music, and how people perceive peak musical experiences. Not surprisingly, much research has been conducted using what are considered to be some of the most important compositions of the Western musical canon, and much research has focused on the performances that most epitomize these pieces.

Only a few researchers have examined how musicians respond to less-than-ideal performances of great works. This discrepancy is easily explained: it is easier to agree on the greatest performances than to sort through the voluminous examples of poor performance. This is due, in part, to a simple reality: great performances are rare. It is exactly this reality; however, that makes research involving poor performance relevant. While the aesthetic experience is what draws (and keeps) people enjoying music, it is not the everyday experience of most musicians, listeners, and teachers. Examining how musicians respond to a performance of variable quality would be a valuable contribution to the growing understanding of discrimination in music.

The purpose of this study was to determine how errors in performance would affect perceptions of performance quality by musicians while listening to an excerpt of Bach's *Bourrée* from the *Suite for Unaccompanied Cello #3*. The importance of this study was to empirically illustrate the response of trained musicians to poor performance, to better understand the musical experience as it unfolds in time, and to attempt to understand and generalize which aspects of performance are most salient from a temporal perspective.

CHAPTER 2
REVIEW OF LITERATURE

The process and speed by which people arrive at decisions has been a topic of much interest to scholars for many years. Entire texts and courses of study have been designed to examine this subject, and major movements within the psychological field have been driven by epistemological arguments advocating a particular approach to decision-making. There are fields of study (heuristics) as well as entire journals (*Journal of Behavioral Decision Making)* devoted to the topic, as well as several rationales that attempt to explain the decision-making process.

Some view the decision-making process as a pattern whereby people make decisions after deliberating over a range of available options. This was the practice of discrimination advocated by numerous thinkers throughout history, from Plato to Benjamin Franklin (who is often credited with the "balance-sheet" style of decision-making). French philosopher René Descartes expressed this viewpoint by saying, "Divide each difficulty into as many parts as is feasible and necessary to resolve it." This perspective is reliant on acquiring a sufficient amount of relevant and correct data on which to base decisions, which is then considered along with the personal perspective of the decision-maker. The combination of these two factors leads to the decision.

Advocates of this sort of decision-making process often deal in terms of *utility*. Utility is neither the outcome of the decision nor the factors that load into the decision, but rather is a variable derived from the observed choices. It has no value in and of itself, and is usually parsed into two types: decision utility and experienced utility. Kahneman & Snell (1992) define decision utility as being: ". . . defined by the sign and weight of a possible outcome in the context of a choice," whereas experienced utility is ". . . defined by the quality and intensity of the hedonic experience associated with that outcome" (pp. 187-188). Because utility is, to some extent, created by the individual, the rationality of an individual's utility is extremely important in guiding a person toward an informed choice.

Recent research, however, suggests that the process by which people make decisions may occur very rapidly, and on levels that the conscious mind does not register. Much of this research was summarized in Malcolm Gladwell's *Blink*. Within this

paradigm, discriminations are more automatic, and can be set into motion by a variety of environmental factors or associations. Galdi, Arcuri, and Gawronski (2008) define these automatic mental associations as " . . . those associations that come to mind unintentionally, that are difficult to control once activated, and that may not necessarily be endorsed at a conscious level" (p. 1100). These associations may differ from consciously held beliefs, which are thoughts and ideas individuals explicitly express or state as accurately reflecting their ideology or decision-making process. Because automatic mental associations may occur at a level that the conscious mind does not register, attempts to examine such associations are often couched in the form of implicit measures.

Implicit learning is ". . . a fundamental characteristic of the cognitive system, enabling the acquisition of highly complex information without awareness" (Reber, 1967). Implicit measures of decision-making can take many forms, but usually rely on "speeded categorization tasks" (Galdi et al., 2008, p. 1100). A prominent and widely-discussed example of such a measure is the Implicit Association Test (IAT). Many forms of the IAT are now available via the internet, from sources like the Social Psychology Network or Project Implicit at Harvard University (Nosek, Banaji, & Greenwald, n.d.). Essentially, an IAT requires that participants pair concepts (i.e. fast and young), with the supposition that the more closely related concepts are, ". . . the easier it is to respond to them as a single unit" (n.p.). Participants should be able to respond faster to concepts which are more closely related (i.e. fast and young) than concepts which are not as strongly associated (i.e. fast and old). An IAT measures the speed of responses to calculate implicit associations.

Implicit Associations may have applicability to music as well. Beyond some of the obvious applications, such as the use of visual cues to make rapid judgments about performers, some scholars have suggested that tonality can function as an implicit system through mere exposure. Tillman, Bharucha, and Bigand (2000) developed self-organizing maps (SOM) that simulated the way that tonality could be learned in this fashion. To test their SOM, the researchers created a learning algorithm with no knowledge of tonality which "adapted to the regularities of harmonic relationships through repeated exposure to tonal material" (p. 907). Based on the performance of this

algorithm, the researchers suggested that knowledge of tonality can be acquired "without specific action or control" (p. 907). IAT research seems to suggest that a variety of factors, many of which may be unconscious, are affecting the manner in which decisions are made in a palpable fashion.

Two such factors that operate to affect decision-making in this fashion are primacy and recency. Primacy is defined as the tendency for greater recall of items at the beginning of a list (as opposed to items in the middle of a list), whereas recency is defined as the tendency for greater recall of items at the end of a list (as opposed to the middle). These factors have been the subject of a variety of investigations in the field of psychology for some time, with documented research in this area occurring as early as 1885 (Ebbinghaus, 1964). Some even characterize the early investigations of Ebbinghaus into primacy and recency as the beginning of experimental psychology (Crowder, 1976). Seashore (1908) characterized primacy and recency as being two of the four secondary laws of habit (primacy, frequency, intensity, and recency), and said, "Other things being equal, the first or primary association will dominate" (p. 123). Early writings by Welch and Burnett (1924), Jenkins and Dallenbach (1927), Thorndike (1927), and Crafts (1932) were primarily concerned with primacy and recency in the context of serial position and memory.

Serial position research deals with how the placement of an item within a list (its serial position) affects recall of that item. Typically, experimenters begin by presenting subjects with certain kinds of items, ranging from lists of words to simple geometric shapes to visible colors. Then subjects attempt to recall the items they saw, or the order in which the items they were presented, or other salient details (dependent on the nature of the study). While psychologists debate over whether serial position is the effect of a linear memory structure or a hierarchical memory device, there is agreement that primacy and recency are substantial factors in serial position. This has led to the establishment of what is known as the "serial position curve" (see figure) which is sometimes referred to as the inverted-U.

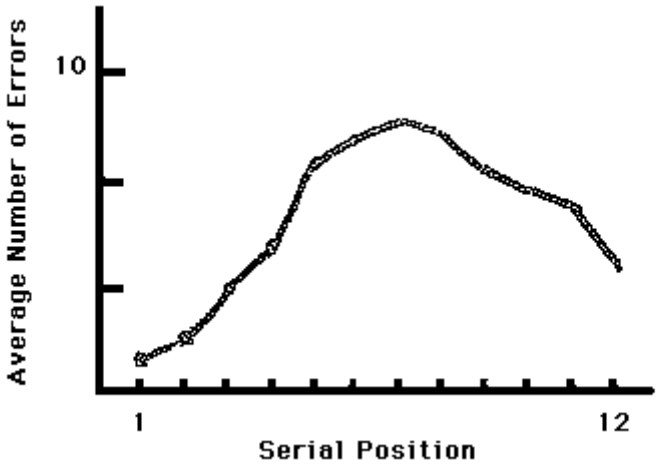

Figure 1. The serial position curve. From McRary, J.W., & Hunter, W.S. (1953). Serial position curves in verbal learning. *Science, 117*(3032), 131-134.

As a construct, the inverted-U has appeared in a variety of scholarship. Within the field of music, for instance, the inverted-U is one explanation for the relationship between complexity, familiarity, and preference. Used in this sense, the U represents the way in which a listeners' preference for a composition is heightened by complexity and a degree of familiarity. Initially, low levels of complexity and familiarity provoke little response. As familiarity and complexity increase, so does preference. This appreciation increases until the music becomes too complex or unfamiliar, at which point the preference of the listener declines. Hargreaves (1984) examined this question in a series of experiments, most of which involved repeated exposure of subjects to musical selections at various intervals followed by post-tests of preference. He found that, for the most part, the changes in the preference of subjects due to repetition did conform to the inverted-U paradigm. Hargreaves' investigations also led him to suggest that perhaps "recency of experience" interacts with familiarity in preference decisions. This illustrates an important distinction: while familiarity can certainly interact with serial position effects, it should not be considered synonymous with recency.

The effect of serial position on memorization of musical materials has been examined from a variety of perspectives, but the findings of various researchers do not suggest that there is agreement on the topic. Williams (1975) completed a variety of analyses of 630 trials of three-, five-, and seven-pitch sequences, and determined that

primacy and recency did play a salient role in pitch recall, although they interacted with several other factors. Specifically, recency interacted strongly with time, such that recall of the last pitch of a sequence dropped each second after the pitch had been sounded. Primacy, on the other hand, seemed to interact most strongly with duration, such that the length of the sequence strongly affected the recall of the first pitch of the sequence.

In another study, Sloboda (1976) examined how subjects would respond to verbal or musical texts that were purposely infused with errors, either of (respectively) spelling or notation. For the purposes of the musical task, he chose to have professional keyboard players sight-read excerpts from compositions by Bendetto Marcello (1686-1739) and Jan Ladislav Dussek (1760-1812). 18 notational errors were introduced into each excerpt (72 total). A third of the errors occurred at the beginning of the phrase, 1/3 at the end of the phrase, and the remainder at other positions. The musicians played the piece twice, and Sloboda inferred that on the second performance the keyboardists would incorporate "memory of higher-order structure" (p. 235). Interestingly, he found that "repetition of a sight-reading task decreases errors on unaltered notes while showing a trend towards increasing errors of altered notes" (p. 234). In addition, he found that the highest degree of inferences during repetition occurred during the middle of musical phrases, and that errors in the middle positions were less likely to be discovered. He hypothesized that this was the result of structural factors.

In a 1987 study Pembrook tried to determine if vocalization would affect the recall process. Subjects were exposed to melodies that they then attempted to recall by singing. At first, his findings seem to be in conflict with other recency findings, because the subjects in his study tended to become less accurate as they sang a melody. Pembrook found that vocalization, however, was a detriment to memorization (as most subjects could not vocalize with accuracy after one hearing), so conclusions based on the practice of vocalizing must be viewed with some skepticism.

Further contradictory findings were the result of a study by Halpern and Bower (1982). The researchers constructed good, bad, and random melodies for examination by their subjects, which included both musicians and non-musicians. Halpern and Bower were interested in whether the quality of the musical material could affect recall in a salient way, and based their study on similar research dealing with the game of chess.

Subjects saw the melodies in question and then attempted to recall them in written form. The musicians displayed superior recall abilities to non-musicians on all of the melodies. However, the musicians also remembered more of the good than the bad melodies, whereas the non-musicians "performed equally poorly on the good and bad melodies" (p. 35). Interestingly, the researchers observed no serial effects for either group no matter the stimulus.

Bergee and Platt (2003) investigated the effect of four distinct variables (time of day, school size, type of event, and performing medium) on solo and small ensemble festival ratings. The researchers drew their data from 7,355 events at a 2001 and a 2002 state solo and ensemble festival. 75 adjudicators took part in the festival, although the festivals did not retain exactly the same adjudicators from year to year. Bergee and Platt demonstrated that time of day, school size, and type of event all displayed statistically important interactions. Based on these findings, at a festival of this type it is better to perform later in the day and to attend a large suburban school. Bergee and McWhirter replicated and extended this research in 2005 using regression analysis and reached similar findings.

Factors affecting Music Preference: How do we know what we like?

Madsen and Prickett (1987) introduce music preference in the following manner:

> Music has a reputation for possessing universal appeal, though a moment's scrutiny of this reassuring idea can raise major questions. Which music? To whom does it appeal? Under what circumstances? The issue of musical preference, its flexibility and its amenity to molding, is of vital interest to us as music educators. (p. 163)

At its most fundamental level, judging music requires making a decision about one's preference for said music. Music preference has been one of the most frequently studied topics in music research, and consequently, a variety of definitions have been applied to this term. Radocy and Boyle (2003) define preference as ". . . an expressed choice of one musical work or style over other available works or styles" (p. 221), while Boyle, Hosterman, and Ramsey state that preferences are ". . . generally recognized as

tangible reflections of attitudes toward music of a particular style" (1981, p. 48). There are comprehensive reviews of the literature in this area (Radocy & Boyle, 2003; Wapnick, 1976), of the measures of preference (Bullock, 1973), as well as general models for music preference (Leblanc, 1980; North & Hargreaves, 1997; Rentfrow & Gosling, 2003); but much of the specific research on music preference has focused on the musical or non-musical factors that affect individual preferences.

Research examining the musical factors affecting preference has been wide ranging, with studies in topics such as tempo, balance, repetition, and performance medium (Geringer & Madsen, 1987; Hargreaves, 1984; Killian, 1987; Kuhn, 1987; Leblanc & Cote, 1983; Wapnick & Rosenquist, 1991; Yarbrough, 1987). Part of the wide examination of musical preference is probably due to its ubiquitous nature: it has applicability to all kinds of music listeners, no matter their age or disposition. Determining how musical elements function for different age groups in preference decisions, for example, has been an area of focus (Leblanc, Sims, Siivola, & Obert, 1996).

Shehan (1985) was interested in the transferability of preference. Specifically, could preference transfer from a taught to an untaught composition if the genre was unfamiliar to the student? Subjects (n=26, 5^{th} grade) received instructions in several genres of music which were unfamiliar to them. In a posttest listening task, Shehan found notable differences in preference for taught and untaught pieces.

Boyle, Hosterman, and Ramsey (1981) attempted to determine what musical characteristics were most consequential in preference decisions among students in grades five, seven, nine, eleven, and college. They found that melody, rhythm, and harmony were the most important factors, but also recognized that the self-report measures used by their study could cause validity issues. This issue (verbal versus operant behavior) is one that is important in preference studies, and many different measures have been taken to try and overcome the problems created therein.

Geringer and Madsen addressed preference with discrimination in a long-range series of studies that, in addition to proposing a hierarchy for musical elements, addressed some of the issues of operant versus verbal behavior as well. This series, that examined how alterations of frequency and tone quality could affect preference and discrimination,

illustrates how subtle alterations in discrete factors can affect music preference. They first (1976) situated this question within trumpet performance, creating a set of excerpts of a trumpeter playing *Twinkle, Twinkle Little Star* with various alterations in tone quality (good, bad). In addition, the instrument accompanying the trumpeter performed a pre-determined intonational condition (sharp, in-tune, flat) each time. The sample (N = 50, music majors) seemed to hold intonation to be more of a factor in preference decisions than tone quality.

In the second study of this type (1981), the researchers wished to replicate the previous study but also ascertain if there was a difference between the operant preference decisions of the subjects and the verbal preferences for flute and oboe duets. Subjects only displayed agreement between verbal and operant behavior at about .60. This highlights again a frequent criticism of musical preference research: individuals do not necessarily report their preferences reliably.

The third study, using vocal excerpts from Schubert and Gounod performed by four soloists, confirmed the findings of the previous studies. Intonation was still the most strongly identified characteristic, but subjects (N = 48, music majors) displayed the ability to discriminate across several categories as well. In the fourth study in the series, subjects (N = 100, music and non-music students) heard excerpts of classical and popular music, some of which had various alterations in frequency and quality. Madsen and Geringer found that subjects generally preferred "sharper" stimuli, and that while music majors could assess frequency more accurately, they could not assess tone quality any better than non-music majors. These studies suggest that musical preference can be altered by the manipulation of pitch, and further, that pitch can play a more decisive factor than tone quality in effecting preference.

Non-musical factors can be just as consequential as musical factors in effecting preference. Non-musical factors can include items such as gender, ethnicity, recording format, setting, socioeconomic group, etc., and may operate on a conscious or non-conscious level. In a study that addressed recording format, Geringer and Dunnigan (2000) examined the effects of several factors on preference: digital versus analog recordings, headphones versus loudspeakers, and performance medium. They prepared recordings of four live concerts of three student ensembles in both digital and analog

formats. Listeners could switch back and forth between the "audio" formats at will to decide which they preferred. The researchers found that "digital" was rated higher than analog in every condition.

A similar study was conducted by Wapnick and Rosenquist (1991). They presented subjects (n=40, undergraduate music majors) with four recordings of piano performances. Three of the recordings were commercial recordings of human performers; the fourth was a sequenced performance recorded from a sampling synthesizer. Subjects in the study seemed to prefer the quality of the sequenced recording to the others, but also rated the tone quality of the sequenced performance lower than the others. Wapnick and Rosenquist concluded that, in certain situations, sequenced performances ". . .might enable musicians lacking the technical skills of concert artists to create recorded performances equal in technical and artistic merit to recorded performances of concert artists" (p. 152).

Killian (1990) examined the effects of performer characteristics on the expressed music preferences of junior-high students. Specifically, Killian was concerned with whether ethnicity or gender would alter subjects' opinions of the performers in a video of *We Are the World*. Most students displayed a preference for performers similar in ethnicity and gender to their own, with males displaying this tendency more strongly than females. In a similar study, McCrary (1993) focused her study on the ethnicity of the listener and the ethnicity of the performer. By using Likert-type scales, McCrary sought to determine if preferences would be affected by ethnicity. She found that black listeners tended to respond more positively to performers that they perceived to be black, whereas white listeners were equally responsive to white or black performers. In this study, race acted as an environmental factor that influenced preference.

Other types of non-musical factors affect preference as well. The effects of suggestion and pre-information as sources of bias in preference have a place in the research literature, even though they have not been examined as thoroughly as have other factors. Studies in this vein sometimes attempt to influence practice by manipulating environmental factors, often to create implicit associations. Two studies have examined the effects of biographical pre-information on preference. In the first, Duerksen (1972) played recordings of two excerpts of a Beethoven piano Sonata for two mixed groups of

music and non-music undergraduate students. Both groups rated the performance in eight dimensions, but the experimental group was informed that one performance they heard was played by a professional musician, while the other performance was played by a student. In fact, both excerpts were performed by the same musician. All of the students in the experimental group rated the student performance lower than the professional performance, and music majors did not appear to discriminate any better than did the non-music majors.

In a similar study, Juchniewicz (2008) presented participants (N = 57, music majors) with biographies for three pianists: a notable performer, a fictitious professor, and a fictitious student. Subjects rated each of the 'performers' even though the stimulus they were presented with contained identical recordings of two excerpts. The student was rated lowest on both excerpts, the professor next highest, and the performer rated highest of all. While these studies do not constitute a large body of research, they seem to suggest that biographical pre-information can bias listener preference in certain controlled scenarios.

Several studies have examined the influence that "suggestion," either from peers or authority figures, can have on preference. Radocy (1976) investigated the role that authority-figure bias could play in classical music preference. He played a variety of classical music excerpts for 20 groups of 15 to 47 students each. He established three groups within his subject pool: no bias, moderate bias (where subjects were given alleged facts about the performers or composers of the excerpts), and strong bias (which utilized the information from the moderate bias condition and value statements). He found that undergraduates were susceptible to bias from authority figures, although they appeared to be more resistant to suggestion with regard to composers and style periods.

In a similar manner, Alpert (1982) examined the influence of various authority figures on preference but added assorted styles of music as a variable. Subjects (N = 82, 4 fifth-grade classes) listened to 30-second excerpts of classical, rock, and country music presented as a market-research type task. Approvals attributed to various authority figures were heard before and after each excerpt. While approvals did have an effect compared to no-approvals on music selection and attitudes, Alpert commented that ". . . the

treatments did not produce the same patterns for each measure. This suggests that musical preferences are multi-dimensional . . .” (p. 183).

Furman and Duke (1988) approached the question of susceptibility to bias from a slightly different perspective by using peers, rather than authority figures, to make suggestions. The researchers played 10 pairs of popular music for participants (N = 160, music and non-music majors) and asked subjects to indicate their preference for each pair. While this process took place, 3 colleagues of the subjects indicated their unanimous approval or disapproval for the excerpts audibly. (These responses were pre-determined.) Non-music majors appeared to be more susceptible to influence, suggesting that training can, to some degree, combat susceptibility to suggestion.

In a related study, Madsen, Geringer, and Wagner (2007) examined the role of conducting styles on the perception of music. Using recordings of the *Blue Danube Waltz* from various Vienna Philharmonic concerts, the researchers constructed a tape that synced performances under five extremely different conductors together as a complete unit. Subjects (N = 108) were divided into four groups: video-only, audio-video, audio-video with cues, and audio only. Participants viewed or heard the performance, depending on their assigned condition, and then rated each performance in seven musical categories. While the subjects in the video condition plainly indicated a change in conductors, almost none of the subjects in the audio condition recognized the nature of the change, instead indicating a high number of changes in the music, some of which did not exist at all. This supported the conclusion suggested by the authors that “ . . . many listeners, regardless of being cued, seemingly ‘hear’ changes in music performances that do not actually occur when asked to indicate any perceived changes . . .” (p. 444).

Factors Affecting Musical Evaluation: Why do we choose one sound over another?

At the heart of discriminatory tasks is the process of reacting to sound, particularly in light of musical expectations. Evaluations of musical performance often operate by making note of when musical expectations are confirmed or violated. In fact, certain aesthetic viewpoints are grounded in this notion. One of the foremost progenitors of this view has been Leonard Meyer, who describes this as a theory of inhibition in his psychologically-influenced *Emotion and Meaning in Music.* Meyer believes that

inhibiting a response to a given stimuli arouses emotion, and that musically, this translates in how listeners perceive successive events in music. In Meyer's words:

> As soon as the unexpected, or for that matter the surprising, is experienced, the listener attempts to fit it into the general system of beliefs relevant to the style of the work. ... three things may happen: (1) the mind may suspend judgment, so to speak, trusting that what follows will clarify the meaning of the unexpected consequent. (2) If no clarification takes place, the mind may reject the whole stimulus and irritation will set in. (3) the expected consequent may be seen as a purposeful blunder. Whether the listener responds in the first or third manner will depend partly on the character of the piece, its mood or designative content. (p. 29)

This process has been investigated in a variety of contexts, and has been shown to occur on a biological level. For example, Maess, Koelsch, Gunter, and Friederici (2001) presented non-musician subjects with several sequences consisting of a succession of in-key chords. Into some sequences, however, they inserted harmonically unexpected chords. Using magnetoencephalography, they determined that such incongruities provoked a response "indicted by early right-anterior negativity," and that this was localized in the area of the brain known as Broca's area (p. 540). Broca's area is a part of the brain also related to auditory language comprehension.

Musical evaluation can take many forms, is a central part of any musical experience, and has always existed in some form within the field of music education. American music education has, for over eighty years, had a form of evaluation through its many contests, festivals, assessments, and adjudicated events. The first contests, originating in the 1920's, were instrumental competitions and awarded trophies or cash prizes to the winners (Rohrer, 2002). These early efforts were often sponsored by music companies, but state organizations soon supplanted commercial interests and assumed control of the contests. As time has passed, music contests have become very pervasive, although many states have adopted a system of ratings (rather than rankings) to avoid some of the issues raised by competition (Williams, 1996).

Rating systems, and similar kinds of musical evaluations, have received much attention from researchers. Researchers have demonstrated that various factors influence the evaluation process in discrete ways. Such research initiatives have examined evaluation in a variety of situations and contexts.

Another component of evaluation is concerned with the creation of valid instruments for performance measurement. Abeles (1973) investigated this question from the perspective of clarinet performance. Abeles considered seven aspects of musical performance in the development of a rating scale: "tone, intonation, interpretation, technique, rhythm, tempo, and general effect (for example, spirit)" (p. 247). Using a review of literature and responses from clarinet players, Abeles developed ninety-four statements that described various aspects of clarinet performance. 50 music teachers then considered two clarinet performances in light of the ninety-four statements, indicating their level of agreement to each statement for each performance via a five-point Likert-type scale. Abeles then used a factor matrix to determine which statements were salient enough to be included in his Clarinet Performance Rating Scale (CPRS). The CPRS was tested under a variety of conditions, and displayed a generally high (.90) degree of reliability.

Another approach to developing a reliable and valid instrument for evaluation was employed by Cooksey (1977), who was concerned with the evaluation in the context of high school choral performance. To determine what criteria are important in that venue, Cooksey examined adjudication sheets, critiques from choral teachers, and essays from choral experts on the subject. This yielded a return of 500 statements that were sorted using categories created by the National Interscholastic Music Activities Commission (NIMAC) into seven areas: balance, diction, intonation, technique, tone, interpretation, and musical effect. After statements that were deemed to be redundant or vague were removed from the list, 147 remained in the initial item pool. In a manner similar to Abeles, the 147 statements were presented to 50 judges with a five-point Likert scale, and employed by the judges in evaluating choral performances. Several analyses were performed on the resultant data, and a Choral Performance Rating Scale (CPRS) was generated based on the results. The CPRS displayed a high degree of inter-judge and

criterion-rated reliability, leading Cooksey to speculate that ". . . some of the difficulties involved in measuring choral performance achievement can be overcome" (p. 113).

The development of reliable and valid instruments for musical evaluation is concerned with creating objectivity in the evaluatory process, but a discussion of objectivity cannot truly take place without considering the perspectives of both the observer/evaluator and the performer/evaluatee, as well as the activity being evaluated. In addressing the former, several studies have considered the conundrum of having one individual serve in both roles. One such study by Yarbrough (1987) was concerned with the usage of several different observation techniques by students in evaluating their own conducting. Previous research (Yarbrough, Wapnick, & Kelly, 1979) had indicated that student self-assessment was not notably different from instructor feedback in evaluating basic conducting skills.

Other scholars have considered the efficacy of self-evaluation in the context of applied music performance evaluation. Bergee (1993) was interested in determining the level of agreement between peers, faculty, and self-evaluation in the context of brass ensemble juries. Using videotapes to present the performances, Bergee asked 10 faculty members at two different universities to evaluate recorded brass jury performances. The first group of faculty (n=5) evaluated 10 students at first and eight students one year later, while the second group of faculty (n=5) evaluated another distinct group of eight students. All of the performers were music majors and either graduate or undergraduate students. In addition to the evaluation by faculty, Bergee also had a group of peers from 'University A' evaluate the performance videotapes from 'University B.' The peers included four graduate brass students, a brass undergraduate, and one brass faculty member. Finally, the performers themselves self-evaluated their performances. Bergee had subjects use a Brass Performance Rating Scale (BPRS), which he had developed and tested in two previous studies, as the evaluation instrument.

Several of Bergee's conclusions are notable. First, he found that scores given to caption ratings appeared to have a relationship, suggesting that perhaps the performances were evaluated on a global level despite the presence of captions. Second, he found that faculty and peer evaluations correlated at a high level (from .86 to .91), while interjudge reliability within the faculty group was also high (from .83 to .89). Third, he noted that

while peer evaluation could be considered to have integrity, it was generally more positive/inflated than faculty evaluation. Finally, he found that self-evaluations did not correlate with peer or faculty evaluations with any degree of strength, but noted that there was not a trend towards inflation such as that found in previous research (Bergee, 1993, p. 24-26).

This investigation was continued by Bergee in 1997 with several modifications. In the second study, Bergee included woodwind, voice, percussion, and string performers (along with brass) as well as faculty members typically called upon to evaluate these areas. The inclusion of these different instrument families meant that the previous measurement instrument (BPRS) was no longer practical, so Bergee used MENC solo adjudication forms in its stead. As with his previous research, Bergee found that correlation between faculty and peer groups was high (.61 to .98) while self-evaluation correlated poorly with other types of evaluation. Faculty interjudge reliability, however, was not as high as Bergee previously found. Bergee conjectured that inadequately sized judging panels, or the experience and longevity of faculty members (both low) could have contributed to this result.

While Bergee examined evaluation on a large scale, Fiske (1975) examined evaluation in the context of judge qualification(s). Subjects, consisting of panels of wind/non-wind and brass/non-brass specialists, listened to audition tapes of 32 trumpeters performing Persichetti's *The Hollow Men*. Participants rated the performances in a variety of categories, but Fiske found that the "overall" rating was related to all other traits, and recommended that this be the only rating employed in tasks of this nature. He also found that the non-wind and non-brass judges could rate the performances, in many respects, as well as their contemporaries.

Fiske (1977) followed this study with an examination of the relationship of several distinct judge characteristics (music performance, adjudication experience, judge performance ability, and judge nonperformance music achievement). Subjects (N = 33, recent music education graduates) again listened to Persichetti's *The Hollow Men* and rated each performance 1 to 5 in five categories: intonation, rhythm, technique, phrasing, and overall. He found that being a good performer does not necessarily make one a good

judge, and interestingly, being accomplished in nonperformance music achievement (e.g. theory and history) actually seemed to make one a worse judge.

In a similar study, Hewitt and Smith (2004) asked subjects (N = 150) to evaluate a recording of six junior high trumpeters using the Woodwind Brass Solo Evaluation Form, a criteria-specific five-point rating scale that examines seven sub-areas. The subjects comprised in-service teachers, upper-division undergrads, and lower-division undergrads, and Hewitt and Smith examined this sample from the perspective of their teaching career-level and primary performance area (brass, non-brass). They found that the instrument of the evaluator didn't appear to matter, nor did the teaching experience of the evaluator (although this conflicted somewhat with Fiske's findings). Some of the teaching experience findings may be attributable to the selection of music evaluated (junior-high trumpeters).

Wapnick, Flowers, Alegant, and Jasinkas (1993) examined the issue of musical evaluation in piano performance by asking subjects to choose between pairs of Listz's *Totentanz* performed by a variety of pianists. They created a set of four different conditions to determine the consistency of their subjects in evaluation (N = 80). One group indicated their preference between each pair, another indicated their preference but assigned a score, another was asked to rate using a 7-point scale, while the final group indicated their preference using score + scale. They found that scores + scales did not improve consistency, and that subjects using scales actually did better without a score. One interesting finding of the study was the extent to which slower tempos seemed to affect consistency negatively.

Wapnick, Ryan, Campbell, Deek, Lemire, and Darrow (2005) used audio examples of pianists at a Van Cliburn competition to determine what effects tempo and duration would have on evaluation. Subjects (N = 167), who were all music students, listened to excerpts of either 20 or 60 seconds, and rated the performance in categories such as accuracy, expressiveness, musicality, etc. They determined that longer excerpts were rated higher than shorter ones, although subjects didn't appear to need longer than a minute to come to consensus. Their results did seem to indicate that the longer excerpt ratings displayed greater reliability.

Not all researchers seem to agree that evaluation of performance is a valid and reliable construct for study, however. Thompson and Williamson (2003) implemented their study of musical evaluation using an instrument from the Royal College of Music (RCM). This instrument, which used a 1-10 point scale, included headings such as "overall quality, perceived instrumental competence, musicality, and communication" (p.24). Three evaluators rated performances given to satisfy recital or exam requirements at the RCM using this instrument. The researchers found that some of their data conflicted with the findings of earlier research. Specifically, they found global ratings to be no more reliable than category ratings, and in general, demonstrated reliability issues amongst all of the ratings. They concluded by saying, "For now, researchers who wish to employ performance assessment as a dependent measure in experimental studies may have to accept that musical performances are simply not open to reliable and constant scrutiny of the type they might wish" (p. 38).

Other researchers have focused on aspects of adjudication that are directly related to appearance, visual stimuli, or labeling. Cassidy and Sims (1996) examined the responses of groups informed of disability versus groups ignorant of disability in evaluations of a choir from a school for children with various disabilities. Participants (N = 209, 119 sixth- and seventh-grade peers & 90 adults) evaluated an 8.5 minute performance video using a choral adjudication sheet that included standard categories (i.e. tone, intonation, diction, etc.), but that also had a few group-specific questions and some open-ended questions. The researchers found that participants without labeling or visual information gave the lowest ratings, and that students rated the group lower regardless of the labels. Adults, however, when provided with the label rated the group higher than all other groups.

Focusing on a different target population, Johnson and Stewart (2005) asked music educators in their sample (N = 201) to assign students to wind instruments based on pictures taken either from a full-face or dental-only perspective. They were interested in ascertaining whether ethnicity or gender affected the instrument chosen. The researchers concluded that neither ethnicity nor gender seemed to have an effect on instrument assignment, although the subjects in the dental-only perspective (n = 98) could have determined the ethnicity of the students from the pictures with careful study.

A notable examination into the effect of gender on evaluation was authored by Goldin and Rouse (2000), who explored the effect that blind auditions have had on female musicians in the orchestral field. Citing demographic data from rosters of nine U.S. orchestras (the big five: New York, Chicago, Cleveland, Boston, and Philadelphia; and Los Angeles, San Francisco, Pittsburgh, and Detroit), the researchers estimated that before the advent of blind auditions less than 10% of the new hires in these orchestras were female, while after their implementation women accounted for 35-50% of all new hires. In addition, by examining audition records from these same ensembles, the authors calculated through regression analysis that the presence of a screen in orchestral auditions accounts for 33% of the probability of a female being hired. While the success rate for all major orchestral auditions is extremely low, the success rate of females in blind auditions of this type is 1.6 times higher than if there were not a screen in use. In this case, the absence of a visual cue makes the evaluation fairer.

In a study that focused more on behaviors, Juchniewicz (2005) examined the effect of physical movement on performance ratings. He asked subjects to rate pianists who exhibited different levels of movement (no movement, head and facial movement, full body movement) during performance. Participants' ratings were positively affected by the movement of the performer, leading Juchniewicz to speculate that ". . . the visual aspects of the performer(s) may be another substantial component of the overall music presentation" (p. 26).

Focusing on conductors rather than performers, Fredrickson, Johnson, and Johnson (1998) prepared video recordings of 20 undergraduate students walking to the podium, picking up a baton, and conducting one measure of common time. Participants (N = 165) evaluated the conducting of the undergraduates using 10-point Likert-type scales. The pre-conducting behaviors of the undergraduates seemed to have an effect on the subjects' perceptions of their conducting ability, with poor pre-conducting behavior affecting perceptions negatively and good pre-conducting behavior affecting perceptions positively.

VanWeelden and McGee pursued a somewhat different line of inquiry in 2007, when they asked participants (N = 353, undergraduate music majors) to rate conductors of various ethnicities. The conductors (2 Caucasian, 2 African-American) were video-

recorded performing two selections: a spiritual and a piece of Western art music. Unbeknownst to the subjects, however, the same ensemble performance was heard as each individual conducted. Subjects viewed the performances and were asked to rate each ensemble and conductor. Subjects rated the Caucasian conductors and their "ensemble" higher for Western art music, and rated the African-American conductors and their "ensemble" higher for the spiritual. The race of the subjects (Caucasian = 101, African-American = 252) was not an important factor in the results.

The findings of the previous studies seem to suggest that visual cues can have an effect on musical evaluation. While this would not constitute a revelatory finding to most, having this information in an empirical format certainly adds weight to the discussion of the influence of environmental factors on evaluation. There is a substantial body of evidence that seems to suggest that evaluation and adjudication can be influenced by implicit associations induced through visual stimuli.

Another study by VanWeelden supports this notion (2002). The researcher examined the role that body-type (endomorphic and ectomorphic) played in the evaluation of conductors. Participants (N = 163) were asked to rate six different choral conductors (3 endomorphic and 3 ectomorphic). Although subjects saw different conductors in each video, the same audio track had been synced to all six conductors. Although two of the ectomorphic conductors in the study were rated differently by subjects, ". . . results of the four-way ANOVA indicated that the conductors' body types did not affect their ratings" (p. 174). These findings seem to conflict with some of the findings of other researchers, most notably with a series of studies by a group of researchers that have examined the effects of attractiveness on the evaluation of performances.

The first, by Wapnick, Darrow, Kovacs, and Dalrymple (1997) addressed this phenomenon in the context of vocal performance. Participants (N = 82) viewed, heard, or viewed + heard individuals singing a classical music excerpt on video. Each performance was seen from several different camera angles. The researchers reported that confounding variables interfered with the ratings for the female singers, but that attractiveness appeared to be a factor in the rating of the male singers. An interesting

ancillary finding of this study was that videotapes generally received higher ratings than did the audio-only condition.

This line of research was first extended by Wapnick, Kovacs, & Darrow (1998) in a study that involved violin performance. For this experiment, 12 violinists performed an audition on video of 2-3 minutes duration. The first half of each video showed the head and neck of the violinist, while the second half showed the upper body (waist up). Participants (N = 72) were once again separated into three conditions (audio, visual, or audiovisual). The researchers found that the attractive students did receive higher ratings than did the unattractive students, but that the attractive students also seemed to play better than the unattractive students. Attempting to resolve this apparent sampling error led the authors to speculate that perhaps the attractiveness bias begins so early in life that it had already led to increased opportunity amongst these individuals, whose mean age was 25 and had an average of 19 years of violin study.

The third study in this series attempted to control for some of the issues of the previous two studies by using 6th grade pianists, each with two years of experience, as the performers. In this study, Wapnick, Mazza, and Darrow (2000) found that their results supported the idea of an attractiveness bias, because while attractive students were rated higher than less attractive students under the audio condition, the difference between these groups under the audiovisual condition was even more pronounced. The nature of the evaluatory process for this entire series, however, most often involved Likert-type scales, and reliability for these measures was not established.

The fourth study in this series (Ryan, Wapnick, Lacaille, & Darrow, 2006) used performances from the Eleventh Van Cliburn competition to create stimulus material. Eighteen of the 30 performers in the competition were viewed by subjects on video performing one-minute excerpts of solo piano performances. Participants (N = 227, all musicians) were assigned to either an audio or audiovisual condition and rated performances in six categories on a 7-point Likert-type scale. In addition to these participants, 38 more individuals were assigned to a video-only group to evaluate the attractiveness of the performers. The researchers found that this study, focusing on performers at the top of the spectrum, conflicted with their previous research in some

ways. The attractiveness bias was not as powerful in this study as it had been in previous studies.

Ryan and Costa-Giomi (2004) sought to evaluate how attractiveness could affect the evaluation of 6th grade pianists. Participants (N = 75, children, undergraduates, and music education majors) viewed videos of 10 performances by young pianists that were each less than a minute in duration, and rated these performers on a scale from 1 to 7. Their results seemed to indicate that there was an attractiveness bias in effect, but that it did not function equally across the entire performer spectrum. For example, being attractive was an asset for the more attractive females, but the less attractive males seemed to receive higher ratings. The authors could not demonstrate high reliability for some of these findings, however.

Music in Time: A Concatenationist View

Evaluations of musical performance occur, informally, all the time. At the end of a concert, a CD, or a song on the radio it is possible to discriminate for or against what one has just heard. This sort of time-sampling raises inconvenient issues, however, for researchers in evaluation and preference. If a listener assesses a performance at its conclusion, what is the listener assessing? Are they assessing the gestalt quality of the experience? Or assessing what most recently occurred?

These questions arise because music is a temporal phenomenon. Visual art is created in time, but as an artifact persists after its creation. It can, by virtue of this fact, be re-examined at multiple points in time. While the perspective of the observer may change, the piece of art itself stays reasonably the same. Music, on the other hand, occurs in time. A musical score is a representation of a piece, but it does not constitute the composition itself. Even recordings, which do not change from one listening experience to the next, are interpretations of a musical entity. Music, unlike visual art, occurs and then is gone.

In some regard, discriminations after the fact are wholly appropriate, and no substitute may be made for them. To consider the form of a musical composition, for instance, it is necessary to recall previous themes and place them in context within the whole. Discrimination of this sort can be very useful, as is evidenced by its pervasiveness in film criticisms, restaurant recommendations, book reviews, and the like.

Post-hoc evaluations of experiences are, however, based on recall of previous events. As such, they have limitations that correspond to issues of memory and cognition. In a way, it is akin to a doctor asking a patient, "How do you feel now?" The patient must make a value judgment, and decide if the doctor wants to know how he felt when the question was posed, or how he feels when he gives his response. If the patient decides to answer based on when the question was asked, he will have to recall that moment in time.

A theory advanced by philosopher Jerrold Levinson (1997) attempts to address this phenomenon in music. Levinson coined the term concatenationism to describe the "moment to moment cogency" of musical experience. He stipulates that four propositions comprise its basic principals (pp. 13-14):

> 1. *Musical understanding* centrally involves neither aural grasp of a large span of music as a whole, nor intellectual grasp of large-scale connections between parts; understanding music is centrally a matter of apprehending individual bits of music and immediate progressions from bit to bit.
> 2. *Musical enjoyment* is had only in the successive parts of a piece of music; and not in the whole as such, or in relationships of parts widely separated in time.
> 3. *Musical form* is centrally a matter of cogency of succession, moment to moment and part to part.
> 4. *Musical value* rests wholly on the impressiveness of individual parts and the cogency of the successions between them, and not on features of large-scale form per se; the worthwhileness of experience of music relates only to the former.

Concatenationism, as Levinson purports it, is primarily a reaction to architectonicism, which he describes as the idea that the height of musical understanding is inextricably connected to the recognition of form. Concatenationism, which Levinson based on the writings of English psychologist Edmund Gurney (1847-1888), maintains that music is different from other arts in this regard. Levinson says, "The parts of an

architectural façade can be taken in more or less in one sweep; the parts of a symphony cannot" (p. 2).

There has been some research that supports Levinson's claim. Bigand (1997) exposed subjects to a variety of melodies, some of which were constructed to resemble others. Even non-musicians demonstrated the ability to group melodies with similar structures together, but upon further examination Bigand found that changing just the first five tones of a long melody could cause subjects to estimate that more than 40% of the pitches in the composition had been changed. The listeners in this study, despite musical training, were sensitive to small changes of structure, often overestimating the effects of those changes.

At the same time, other alterations in musical factors have been shown to have little effect on preference, particularly alterations in global structure. Researchers have reordered movements of Beethoven sonatas (Konecni, 1984) and the Bach *Goldberg Variations* (Gotlieb & Konecni, 1985), and even rearranged sections of the first movement of Mozart's *Symphony in G Minor (K550)* (Karno & Konecni, 1992) but found that such alterations had little effect on participants' subjective judgments when compared to the original. (Although it is important to note that the original versions were preferred slightly in each case.)

Results such as those found in the previous two paragraphs led Tillman and Bigand (2004) to conjecture that listeners, trained or not, could perceive small-scale alterations in music conforming to a Western tonal style. They hypothesized that this was possible because of implicit knowledge of tonality, which enabled listeners to make discriminations between short musical ideas (between 20 seconds and three minutes). At the same time, Tillman and Bigand suggested that the perception of large-scale musical structures were more difficult for subjects, and that, "Global structures seem to have weak influences on perception, and local structures seem to be much more important" (p. 218).

Concatenationism, as Levinson explains it, marginalizes the act of form comprehension in the musical experience. Levinson's theories have primarily been answered by Peter Kivy, another philosopher with an emphasis in aesthetics. Kivy

maintains that understanding of form is endemic to musical understanding, and that the centralization of one's perceptions in this format is indispensable.

Temporal Measurement of Musical Experiences

> A creature cannot be beautiful if it is too great, for contemplation of it cannot be a single experience, and it is not possible to derive a sense of unity and wholeness from our perception of it.
>
> -Aristotle, *Poetics*

There are a variety of ways to measure musical preference, ranging from static, verbal measures to continuous, operant measures. Much research has been attempted using the former, primarily for reasons of convenience. In the case of a static measure, subjects are asked at a specific moment in time to record their preference for a recording, performance, etc. The response can take the form of a Likert-type scale, a semantic differential, a nominal descriptor (good, bad), or whatever form the researcher desires. In cases where a gestalt-type rating is desired, a static measure at the conclusion of the musical work may be the best available choice, particularly in light of the findings of Bergee (1993) and Fiske (1975, 1977).

Static measurement may not adequately capture the entirety of the listening experience, however. Oftentimes, music is structured such that it will elicit the most appropriate response only at its conclusion. As such, it seems reasonable to hypothesize that the way that one feels about a musical performance can be influenced by the proximity, or recency, of the last "big moment." In fact, a variety of studies outside of the field of music have shown that the final moment of an experience can have tremendous impact on how subjects perceive an experience. For instance, Kahneman, Fredrickson, Schreiber, and Redelmeier (1993) exposed subjects to an unpleasant experience: immersing their limbs in frigid water of varying degrees. In the first trial, subjects immersed their limbs for 60 seconds at 14 degrees Celsius, while in the second trial, subjects immersed their limbs in water for 60 seconds, but then kept their limb in the water for 30 additional seconds while the temperature of the water was raised to 15 degrees Celsius, which was ". . . still painful but decidedly less so for most subjects" (p. 401). When subjects were asked to choose a trial to repeat, the majority chose to repeat

the second, longer trial. This led the researchers to hypothesize that the intensity of the negative experience was a stronger factor than the duration of the trial. In the case of the second trial, it is apparent that the static measure employed at the end of the trial yielded considerably different results than would a continuous measure. Despite the unpleasantness of the water, subjects were ready to repeat the second trial simply because the ending was more pleasant.

This phenomenon has also been explored by researchers in social psychology, particularly with regard to the length and strength of unpleasant experiences. Kahneman and Fredrickson (1993) were interested in how the duration of an activity would affect subjects' perceptions of that activity, especially when that activity was of very high intensity or strength. The researchers showed film clips to subjects; some of the clips were considered aversive, while others were considered pleasant. All of the clips were of varying duration. The first group of subjects (n=32) rated each clip in real time (continuously) and at its conclusion (static). The second group of subjects (n=96) viewed the same clips, but ranked them later based on recall/memory. Kahneman and Fredrickson found that increasing the duration of a pleasant clip did not increase the rating of the clip in either measurement format. Increasing the duration of an unpleasant clip, on the other hand, caused ratings to become more negative (in either format).

A musical application of this concept can be found in a study by Rozin, Rozin, and Goldberg (2004). The researchers had subjects push a pressure-sensitive button while listening to various musical selections, and then later had subjects report the "remembered affective intensity" of each selection. The analysis of Rozin et al. determined that the recalled affect bore little relationship to the continuous ratings, and that the rating of the peak experience of the listener most closely approximated the "remembered affective intensity." Rozin et al. said, "Listeners rely on the peak of affective intensity during a selection, the last moment, and moments that are more emotionally intense than immediately previous moments to determine post-performance ratings" (p. 15).

Static measurements also have obvious limitations, however, as they can fail to account for the range of feelings that transpire during a musical experience. This issue created by this quandary is one of temporal assessment. To adequately capture some of

the nuances of musical experience, it is necessary to "sample" the changes in a listeners' perception across time, not merely at one discrete point. As stated by Madsen and Fredrickson (1993), "Every musician realizes that music moves through time, yet many of our responses to music come after the fact, when the music is over" (p. 47). To put it another way, while snapshots of any experience are useful, they do not capture experience like a motion picture.

To address these concerns, researchers focused on computer-assisted assessment. One of the primary instruments employed for this purpose is the Continuous Response Digital Interface (CRDI). The CRDI is a dial (potentiometer) which, when turned by a subject, can be used to express a subjects' reaction to music. The dial gathers measurement in the form of voltage fluctuations, which is then translated via computer into digital information and then into a numerical format. The largest advantage to the CRDI is its ability to take measurements across time, such that subjects have the opportunity to respond to minute changes in music as they perceive them, and to express their response to these changes without the necessity of words or other communication systems. The CRDI can be set to take time-samples at minute intervals (up to 10,000 times a second, but most commonly twice a second) and records data along a continuum, which can be adjusted anywhere from 0 to 255 depending on the needs of the investigation. The CRDI has also been demonstrated to be a reliable instrument in a variety of situations, displaying test-retest reliability ranging from .64 to .90 in various studies (Gregory, 1996).

Applications of the CRDI have been broad in scope, and a cursory examination of some of these uses provides context for the application of this measurement tool. The brief summary that follows should not be considered a comprehensive list of CRDI research. Rather, the studies found herein are included because they are thought to be germane to this investigation in some fashion.

One of the earliest uses of the CRDI was in the global evaluation of choral performances. In this study, Robinson (1988) had subjects manipulate the CRDI dial along a negative-positive continuum, essentially casting a Likert-type task in a temporal format. In a study designed to address some of the issues proposed by Robinson, Brittin (1991) compared subjects' static preferences for musical excerpts to measurements

gathered with the CRDI. She found that CRDI responses tended to be more positive than static measures.

In a follow-up study to the 1990 study, Brittin and Sheldon (1995) expanded on the previous research in two ways. First, the subject pool was expanded from non-music majors to music majors (n = 100) and non-music majors (n = 100). Second, the stimulus material was changed from popular music to art music, and encompassed string, instrumental and wind music from different musical periods. Half of each group rated the music at its conclusion using a 10-point Likert-scale, while the other half rated the performance using the CRDI "to continuously signal their degree of liking throughout each excerpt" (p. 38). The researchers found that music majors' static responses tended to correspond with the means of their continuous responses, while non-music majors' continuous responses were consistently higher than their static responses.

Colwell's 1995 study is related to the work of Robinson, Brittin, and Brittin and Sheldon, in that it used the CRDI as a global evaluative instrument but also compared CRDI responses with other self-evaluation tools. Subjects taught either children or peers and then were evaluated using either CRDI or a behavioral checklist in the category of teacher intensity. The expert educators consulted for the study ". . . reliably evaluated teaching as effective or ineffective using the CRDI" (p. 18), but these findings only moderately correlated with the behavioral checklists completed by the investigator. Colwell offered several rationales for this discrepancy, among them the possibility that the ". . . criteria for high-intensity teaching were not the same" (p. 19).

In some of the previous studies, the mean rating of a continuous measure was used for the purposes of comparison. Using means in this fashion is only appropriate in specific circumstances, because summative information of this type can be misleading in the wrong context. Duke, Brittin, and Colprit illustrated this point in two studies (1997 & 2001). In the first of the two (Brittin & Duke, 1997), subjects were music or non-music majors (n = 40), who heard nine musical excerpts (none of which were longer than 1 minute 16 seconds) from major orchestral works. The subjects listened to each excerpt twice and were asked to indicate their perception of the musical intensity of the excerpt. Listeners were assigned to one of five different response groups, each of which allowed for static/continuous ratings in a different fashion. For instance, listeners in the

summative/continuous group were instructed to rate the excerpt at its conclusion the first time, and then rate the excerpt throughout its performance the second time. Brittin and Duke found that the summative/static responses were consistently higher than the continuous means.

In the second study of this type, Duke and Colprit (2001) replicated the previous study, but had subjects click a mouse on a computer screen, rather than use a CRDI. The mouse was clicked on any one of 10 buttons representing ordinal categories of intensity. The findings were identical to that of their previous study. Brittin and Duke advised that the use of continuous or static measurement must be determined by the research question, but cautioned against the use of continuous means as a sort of "psychological average," because continuous means do not appear to equate to post-hoc, static measures.

Another of the earliest applications of the CRDI was found by Rentz (1992), who used a vertical form of the CRDI device in a focus of attention task. Subjects listened to an excerpt from *Billy the Kid* by Copland and used the CRDI to indicate which instrument family (brass, percussion, woodwinds, strings, all) became their focus-of-attention using a sliding lever (rather than a dial). Then, Rentz examined the amount of time each subject spent focused on each category. Results showed that non-musicians spent more time focused on brass and percussion, whereas musicians spent more time focused on strings or on three or more families of instruments.

In addition to adjudication/evaluation and focus-of-attention research, the CRDI has also been used to examine aesthetic response to music. The earliest instance of the use of the CRDI in this sense was conducted by Madsen, Brittin, and Capperella-Sheldon (1993). In this study, the investigators sought to determine how subjects would respond to an excerpt from Act I of Puccini's *La Boheme* across time. The resultant data was charted graphically, producing a representation of each subjects' aesthetic experience as it occurred. Every subject in the study indicated that they had an aesthetic experience, and every subject indicated that the CRDI graph approximated their aesthetic experience. In addition, the peak experiences of all of the subjects seemed to cluster at specific points in the music, and this temporal event seemed to be short (15 seconds) in duration.

A follow-up to this study was conducted by Madsen in 1997. Madsen used the same Puccini excerpt but asked subjects to indicate their focus of attention in terms of

five musical elements: melody, rhythm, timbre, dynamics, and everything. Subjects (n = 100) were divided into two groups. Half of the subjects could use a CRDI dial to choose between which of the five elements were most salient at different times during the recording. The other half had only one element on their dial and were instructed to indicate their degree of attention to that specific element. Madsen found that the subjects who chose between elements found dynamics to be most compelling, followed by everything, melody, rhythm, and timbre. When subjects attended to only one element, however, melody was most important, and also most closely related to the aesthetic experiences described in prior research.

Closely related to the aesthetic experience is what is known as "musical tension," which is described by Nielsen (1987) as the quality of tension in the music, and the actual feeling of tension in the listener. In a study designed to replicate Nielsen (1987), Madsen and Fredrickson (1993) asked subjects to listen to the same recording of Hadyn's *Symphony No. 104* and respond to their perception of the musical tension of the piece using a CRDI dial. An overlay placed on the dial represented tension graphically through a series of nine visual stimuli, which progressed from white to moderately dark shading. Nielsen's original study used a pair of "tension-tongs" in place of the CRDI, which were essentially tongs equipped with a potentiometer that could be squeezed to indicate musical tension. The graphs of the tension experience produced by Madsen and Fredrickson's replication closely approximate Nielsen's graphs, particularly with regard to the relative strength and weakness of the "hills and valleys" of the aesthetic experience.

Other researchers have used the CRDI as a discrimination/perception tool, examining error detection, loudness, rubato, tempo, etc. In one of the early studies of this type, Sheldon (1994) asked subjects (musicians and non-musicians) to respond to alterations in tempo (accelerations, decelerations) using the CRDI dial. Subjects were exposed to tempo alteration in one of three conditions: audio-only (listening), audio-visual (listening while watching a conductor), or audio-motoric (listening while moving in some way). She found that musicians more accurately detected alterations than did non-musicians, and in addition, found that subjects were better at detecting acceleration than deceleration, which confirmed some of the findings of previous (non-CRDI)

research (Geringer & Madsen, 1984; Madsen, Duke, & Geringer, 1986; Yarbrough, 1987). In a similar study, Geringer (1995) investigated perception of loudness levels by having subjects listen to ten excerpts, each of which contained at least one prominent crescendo and one prominent decrescendo. The CRDI allowed the researcher to examine the magnitude of perceived change as it occurred in time, and also allowed for a wider range of responses: "A second difference between older studies and more recent studies concerns the nature of the dependent measure, that is, the magnitude estimation of isolated acoustical events versus continuous responses to ongoing music. . . . subjects in this study used a wider numerical range of responses to stimulus change than subjects typically use in the method of 'free' magnitude estimation" (p. 33).

The previous studies usually were concerned with gathering information along one continuum, but Gregory's (1994) examination of preference had several lines of inquiry to measure. Two of the most salient questions from this study were: what music do students prefer? And, to what extent does their knowledge of this music (familiarity) influence their preference? To address these questions in tandem, Gregory had 30% of her subjects (which encompassed an age range from 6th grade to undergraduate) manipulate two CRDI dials simultaneously. One dial consisted of a conventional positive-negative continuum, and was used to determine student preference. The other dial had an overlay with eight possible descriptors, ranging from "totally unknown" to "have performed/analyzed or taught it" (p. 334). Interestingly, one of Gregory's ancillary findings seemed to confirm the findings of Hargreaves (1984). Specifically, when Gregory examined the relationship between knowledge and preference she found little correlation, except among sixth-grade subjects, who displayed the only significant, positive correlation. Gregory conjectured that this supported Hargreaves because the excerpt that subjects were most familiar with was taken from the *Silver-Burdett* music series, and that sixth graders would be most affected by a "recency of knowing."

Another preference study was conducted by Plack (2006), in which the investigator examined the effect of performance medium and listener characteristics on a preference task. Plack used five separate recordings of Giacomo Puccini's *Nessun Dorma* from *Turandot* as stimulus material. Participants heard the original version (with orchestra) as well as arrangements for piano, wind ensemble, marching band, and popular

dance music. Subjects (N=143) were assigned to five groups as well: voice, marching band, wind ensemble, non-musicians, and piano. Participants manipulated a CRDI dial throughout each recording, after being instructed to use the device to indicate their emotional response. Plack found a strong relationship between the music majors and their corresponding performance area, as well as a strong relationship between the non-music majors and their corresponding groups (popular dance, and marching band).

In a study that combined elements of aesthetic research with preference, Southall (2002) investigated how "purposeful distractors" would affect the preference of subjects listening to an excerpt from Puccini's *La Boheme*. Southall created a recording that superimposed one of two audio distractors onto the performance: either telephone sounds (busy signal and ringing) or pink noise. He assigned subjects (N=96) into one of three groups: control, telephone, or pink noise. Each of the subjects responded to the recording by manipulating a CRDI dial during the performance. Southall found that while subjects did respond to the distraction, they recovered quickly from it and seemed to still have an aesthetic experience over the course of the excerpt.

Summary

Discriminating between choices is one of the most valuable, sought after, and studied parts of human existence. While scholars do not always agree on the mechanism by which choices are made, there is a growing body of research that suggests that the decision-making process may work extremely quickly and be influenced by environmental factors. Certain environmental factors, such as serial effects and familiarity, have been shown to palpably influence discrimination in specific contexts.

Preference for music is, at its fundamental level, a discriminatory task. It can also be tremendously influenced by a variety of factors, both musical and non-musical. Evaluation of music performance is both a discriminatory task and a necessary skill for any musician. Prior research has shown that musical training can be a factor in evaluation, and that rating systems for evaluation are often complex instruments.

Computer assisted assessment has made it possible to examine how perceptions of listeners change across time, greatly expanding the possibilities for research. Entire lines of inquiry into the aspects of music that comprise the aesthetic experience have been conducted, and by focusing on the most exemplary models of musical performance, the

results of this research have contributed to a greater understanding of the human condition. These investigations ignore a salient point, however: the greatest performances are also the most seldom heard.

Given this reality, examining how less-than-exemplary performances affect aesthetic experiences would seem to be of the utmost pertinence. In particular, an examination of how variability in performance quality affects listener perceptions across time would represent an important contribution to the field. The present investigation was designed to examine how changes in quality of performance would affect listener discriminations of quality temporally.

Statement of Problem

The aesthetic experience is, arguably, the *sine qua non* of all music studies. It can be reasonably stated that the prevalence of music in every culture is due, in some part, to this phenomenon. As such, the aesthetic experience has been the subject of much speculation and investigation by philosophers, scholars, teacher, performers, and researchers.

Particular interest, of late, has been given to how the aesthetic experience unfolds in time. Descriptions of aesthetic experiences, such as those suggested by Langer's "morphology of emotions" (1957), have been augmented by continuous measurements of the aesthetic experience. Careful descriptions of the "peaks" and "valleys" of particular aesthetic experiences are now possible, as is the examination of the travel-process to these places in time. Despite this focus, it can be argued that the aesthetic experience is not the common musical experience of most people. In fact, it is the elusiveness of this phenomenon, to some degree, that makes it so compelling. The day-to-day experience of most musicians and listeners is less divine and more banal.

Few have attempted to examine less-than-ideal performances, primarily because finding a standard for these performances is more elusive. In fact, bad performances are so ubiquitous that they defy categorization. It is hard to say what makes a performance bad, however, most musicians would agree that they know it when they hear it.

Given that reality, it would seem that a study that examined how musicians respond to bad performance would be appropriate, necessary, and a valid contribution to the understanding of the musical experience. Innovations in technology, created for the

purpose of making poor performances sound better, could be applied in the reverse to make a good performance sound poor. The purpose of this study was to determine how errors in performance would affect perceptions of performance quality by musicians while listening to an excerpt of Bach's *Bourrée* from the *Suite for Unaccompanied Cello #3*. The importance of this study was to empirically illustrate the response of trained musicians to poor performance, to better understand the musical experience as it unfolds in time, and to attempt to understand and generalize which aspects of performance are most salient from a temporal perspective.

Pilot Study

A pilot study was conducted to test the feasibility of this investigation. The most problematic element of the current investigation involved the creation of the independent variable, in this case, a musical stimulus derived from an extant recording. Initially, the *Sarabande* from the *Suite for Unaccompanied Cello #5* (DIDC 20082) was considered as the stimulus recording. This piece was originally proposed as a possible extant recording because it was thought that: 1. slower music would be easier to manipulate, 2. a recording a solo performer would be easier to manipulate, and 3. a performance by an unfretted string instrument would more naturally accommodate changes in pitch.

Once this recording was obtained, it had to be manipulated such that it appeared to exhibit performance errors. The performance was altered by a professional audio engineer using the computer application *Pro Tools 8*, and the original audio track was converted to AAC format. Initial modifications were suggested by the researcher based on evaluation of the harmonic content of the excerpt.

The initial suggestions made to the engineer included altering the following pitches:

Table 2.1 Alterations to Pilot Recording #1

B at 8-9 Second Mark (Measure 2, Beat 3)
C at 17-18 Second Mark (Measure 4, Beat 3)
E-Flat at 32-33-34 Second Mark (Measure 8)
F at 1:31-32 Second Mark (Measure 12, Beat 3)
G at 1:46 Second Mark (Measure 16, Beat 1)
C at 2:04-05 Second Mark (Measure 20, Beat 1)

These suggestions were submitted with a written description of the type of performance that it was hoped the alterations would approximate. In this instance, a performance wherein the cellist was having trouble with lower, sustained notes, but "improved" his performance when repeating the phrase. For this reason, it was suggested that the engineer only alter pitches on the 1st time through each section. It was also

suggested that pitches be adjusted sharper, and that the alterations be close to the range of ½ step.

The initial modifications were accommodated by the engineer where possible. The engineer noted that longer notes were more difficult to manipulate, particularly with "the resonating tones as the player changes strings" (B.C. Beard, personal communication, February 3, 2009). This caused one of the suggestions submitted by the researcher to be discarded (the G at 1:46) because another note was being held into it. In general, the engineer commented that it would be easier to manipulate the recording if the performance were more detached.

This initial stimulus recording was played for several music experts (n = 4) in a single-shot, case-study format. Experts were not told the nature of the investigation, but were asked to listen to the recording and then provide free-response comments on the quality of the performance. Experts agreed that the quality of the performance was generally quite high, but when asked to speculate as to the flaws in the performance provided highly divergent answers. For example, one expert commented that it sounded like the performer was playing "much more detached on the repeat than on the 1st time through," while another commented that the recording was two different performers sliced into one excerpt.

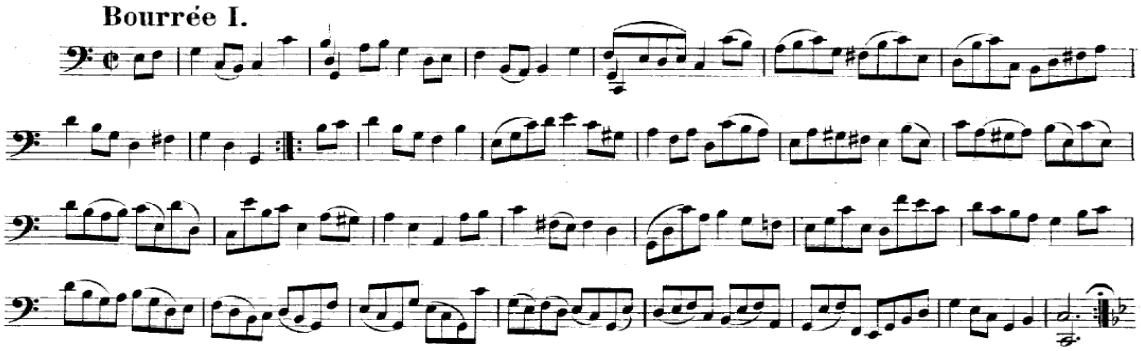

Figure 2. Excerpt from *Bourrée* from the *Suite for Unaccompanied Cello #3* by Johann Sebastian Bach. Public Domain.

To accommodate the suggestions of the engineer and the experts, a new extant recording was obtained. A performance of Yo-Yo Ma performing a *Bourrée* from the

Suite for Unaccompanied Cello #3 (DIDC 20081) was chosen in lieu of the earlier example. (Rationale for this selection, its length, and other pertinent information can be found in Chapter 3).

Using the same method as the previous example, the researcher suggested the following alterations:

Table 2.2 Alterations to Pilot Recording #2

0.1 - C to C# - Measure 1, Beat 4

0.3 - G to G# - Measure 2, Beat 3

0.6 - C to C# - Measure 4, Beat 3

0.9 - D to D# - Measure 7, Beat 1

0.11/12 - D to Db and G to G# - Measure 8, Beats 2 & 3 (respectively)

0.24 - D to D# - Measure 9, Beat 1

0.27 - E to Eb - Measure 10, Beat 3

0.29/30 - E to F - Measure 12, Beat 3

0.34/35 - E to Eb and A to Bb - Measure 16, Beats 2 & 3 (respectively)

0.37 - B to Bb - Measure 18, Beat 3

0.40 - G to G# - Measure 20, Beat 3

0.44 - C to C# - Measure 23, the "and" of 4

0.49 - G to Gb - Measure 27, Beat 1

0.51 - C to C# - Upper note, on the recording-the second note of the double-stop.

The excerpt is constructed of two phrases, and each of the alterations found above occurs during the 1st time through each phrase. To control for performer-created differences between the 1st and 2nd repeats, the engineer was asked to use only the 1st time through each repeat as source material. Subjects would then hear the 1st repeat-altered, followed by the 1st repeat-unaltered, etc.

The engineer was able to accommodate all of the above suggestions except two (G to G#, meas. 8; C to C# at 0.51), which were omitted because of sustained tones

which prevented the edits. The engineer also inserted the last octave jump from the 2nd repeat at the end of the B phrase, because it seemed to contribute to a feeling of finality at the conclusion of the excerpt.

This new version of the excerpt was played for several musical experts (n = 6) in an individual setting. The experts were given the following instructions:

1._____ Rate the quality of the performance from 1 (awful) to 7 (great)
2._____ when prompted. At each prompt, you'll be assigning a number that
3._____ encapsulates how you feel about the performance at that moment.
4._____
5._____
6._____
7._____
8._____

Each expert was prompted at different times during the recording, to assess if the recording was provoking response changes adequately. Experts also submitted free-response information at the conclusion of each listening section. Results indicated that experts felt that the simple Likert-type scale responses somewhat-adequately represented their feelings at a particular moment in time. Experts did not answer uniformly when asked if the Likert response they submitted was a reaction to the music at that moment, or based on recall of previous events. No expert recognized that the performance had been digitally altered, and when informed of the alteration, several commented that the edits were excellent and unidentifiable as such.

Based on the results of the two previous studies, 5 excerpts were prepared following the procedure above. Each excerpt was considered a different experimental condition, and was evaluated by a panel of experts. (See Chapter 3 for a full account of this process). The five experimental conditions were then randomly ordered, and evaluated by a different panel of experts (n=8) using a CRDI dial. The CRDI served as the dependent measure, recording numerical ratings across time. Experts were tested in groups of 4. Instructions to the experts were:

> In a moment, you will hear five performances of a work for solo cello. As you listen to each performance, please turn the dial in front of you to

indicate your preference for each performance. Before beginning each listening task, make sure your dial is in the neutral position. There will be a short break (14 seconds) between each performance during which music unrelated to the study will play. Use this time to return your dial to the neutral position.

At the conclusion of the listening task, the experts completed free-response evaluations and then offered suggestions to refine the purpose of the study. Suggestions included making instructions more specific and isolating subjects (visually) from each other. All suggestions were implemented in future iterations of the experiment. The experts also evaluated their CRDI numerical data, and indicated that it confirmed their rating(s) of the excerpts across time. Examination of individual graphs of the experts' responses indicated that some subjects moved the dial continuously, while others moved it only sporadically. The full range of the CRDI (0 to 255) was used by several of the subjects. The following study was structured based on the responses to the pilot.

CHAPTER 3
METHOD

The purpose of this study was to determine how errors in performance would affect musicians' perceptions of performance quality while listening to an excerpt of Bach's *Bourrée* from the *Suite for Unaccompanied Cello #3* (DIDC 20081). Specifically, the study sought to determine how performance errors within sections of a composition interact with and affect perceptions of subsequent sections. And, to what degree listeners 'recover' from errors in performance when they evaluate sections that follow sections with errors.

Participants

Participants were 129 music students at a large university in the Southeastern United States. Participants were undergraduates (n = 89) and graduate students (n = 40) with an average age of 23.7 years old. The mean years of musical training for the participants was 12, and participants' performance areas included vocal (n = 34), string (n = 11), wind/percussion (n = 76), and piano (n = 9). Participants were recruited for the study haphazardly from the general population of the school of music at the university. Pre-arranged times and sign-ups were encouraged for participation, but some participants were "walk-in" participants who did not arrive at a pre-arranged time.

Musical Stimuli

The musical examples for this study were based on the *Bourrée* from Johann Sebastian Bach's *Suite for Unaccompanied Cello #3*. Only the first 28 measures of this movement (a section sometimes referred to as Bourrée I) were utilized. The first complete measure was considered measure 1. The excerpt consists of two phrases, each repeated once. Phrase A consists of the section from the beginning to the third beat of measure 8, while the B phrase consists of the section from the fourth beat of measure 8 to the third beat of measure 28 (see figure 2). The excerpt begins and ends in C Major, but the first phrase ends on a half-cadence which is followed by a brief tonicization of A minor. A borrowed D7 quickly returns the excerpt to G, which is firmly established as the dominant (V) chord before the excerpt culminates via a perfect-authentic cadence in C.

The Bach *Cello Suite* was chosen as source material after extreme deliberations. First, it was necessary for the stimulus to be of a solo performer, due to restrictions

imposed by the audio editing required for the study. Second, it was necessary that the performance be of the utmost musical and aesthetic quality, but reasonably brief and amenable to parsing without loss of aesthetic quality and continuity. Third, it was necessary that the recording be somewhat familiar to the musicians in the sample, to most closely replicate an evaluatory task (where it is usually assumed that some of the judges will have knowledge of the score being realized.) Finally, it was necessary that the recording be considered seminal, so that unaltered sections of the recording could be considered to be (to most) definitively of the highest quality.

The study used a single recording which was a 1983 performance by Yo-Yo Ma performing the Bach *Suites for Unaccompanied Cello*, recorded on compact disc. This performance was evaluated by a panel of experts, comprising a total of 49 years of experience, and rated using a seven point Likert-type scale. On the basis of this evaluation, the original performance was used for the "good" segments in all subsequent experimental conditions. The total time for the stimulus was 1 minute, 22 seconds.

Stimulus Creation

Sections of the original recording were altered to create two alternate versions of each phrase: a "high-error" and "low-error" segment. These segments, coupled with the unaltered version of the recording, resulted in three versions of each phrase. To create the "low-error" and "high-error" segments, individual pitches within the performance were digitally altered. Specifically, modifications were made to the correct pitches in the original performance to make it appear that the performer was playing wrong notes. In most all cases, pitch was adjusted sharp, because this was thought to more closely resemble performance practice as documented in previous research (Madsen, Geringer, & Heller, 1991). In a few instances, however, adjusting pitch sharp would have resulted in a somewhat acceptable non-chord tone, so in these situations pitch was adjusted flat. The adjustment was typically 12% or more, resulting in an almost a half-step alteration of each altered pitch. The "high-error" segment was designed to present the listener with an obvious mistake almost every three seconds, while the "low-error" segment reduced this rate to only 2-3 mistakes per phrase. Specific alterations are found in Table 3.1. The location in time of each alteration is provided for the repeat where appropriate.

Table 3.1 – Alterations to Original Recording, High-Error.

High Error Version – Phrase A

Time (Min/Sec)	Alteration	Location (Measure, Beat)
0.1	C to C#	Meas. 1, Beat 4
0.3	G to G#	Meas. 2, Beat 3
0.6	C to C#	Meas. 4, Beat 3
0.9	D to D#	Meas. 7, Beat 1
0.11/12	D to Db	Meas. 8, Beats 2

High Error Version – Phrase B

Time (Min/Sec)	Rpt.	Alteration	Location (Meas, Beat)
0.24	0.52	D to D#	Meas. 9, Beat 1
0.27	0.55	E to Eb	Meas. 10, Beat 3
0.29/30	0.57/58	E to F	Meas. 12, Beat 3
0.34/35	1.02/03	E to Eb & A to Bb	Meas. 16, Beats 2 & 3
0.37	1.05	B to Bb	Meas. 18, Beat 3
0.40	1.08	G to G#	Meas. 20, Beat 3
0.44	1.12	C to C#	Meas. 23, "&" of 4
0.49	1.17	G to Gb	Meas. 27, Beat 1

Table 3.2 – Alterations to Original Recording, Low-Error.

Low Error Version – Phrase A

Time (Min/Sec)	Rpt.	Alteration	Location (Measure, Beat)
0.3	0.18	G to G#	Meas. 2, Beat 3
0.9	0.21	D to D#	Meas. 7, Beat 1

Low Error Version – Phrase B

Time (Min/Sec)	Rpt.	Alteration	Location (Measure, Beat)
0.27	0.55	E to Eb	Meas. 10, Beat 3
0.40	1.08	G to G#	Meas. 20, Beat 3
0.49	1.17	G to Gb	Meas. 27, Beat 1

To help control for performer-created variances between the 1st and 2nd repeat of each phrase, only the 1st time through each phrase was used for the excerpts. In cases where a "good" version of a phrase was required during the 2nd repeat, the 1st time was inserted instead of the performer's 2nd time. The very last octave jump (from Phrase B - 2nd time) was inserted at the end of the excerpt, however, because it seemed to end the excerpt more effectively and create the illusion that the performance was complete/seamless.

Modifications were made using the application Pro Tools HD-8 (2008). Each excerpt was controlled for input/output levels. The excerpts were also identical in length (1 minute, 22 seconds).

A true counter-balanced design, making use of all of the three versions of each phrase in each order, was rejected for two reasons. One, it would make the duration of the listening task much longer than advisable (based on previous research). Two, it would result in experimental conditions that lacked connection with feasible performance scenarios. For example, it is unlikely that a performer would play Phrase A exceedingly well the first time and then fraught with errors during its repeat.

In order to create experimental conditions that were more realistic, five excerpts were fabricated using the "good," "low-error," and "high-error" segments. The corresponding rationale for each of these excerpts was evaluated by a panel of experts before the excerpt was chosen for inclusion in the study. Even though a counter-balanced design was not appropriate, the structure adopted (Table 3.3) allowed subjects to "hear" a good performance through rotation of various segments.

Table 3.3 – Phrase Structure of Five Experimental Conditions

Excerpt	Phrase A	Phrase A (Repeat)	Phrase B	Phrase B (Repeat)
1	low-error	good	good	good
2	good	good	good	low-error
3	good	good	high error	low-error
4	good	low-error	low-error	high-error
5	high-error	low-error	good	good

Excerpt 1 was designed to resemble a generally good performance marred by a few mistakes at the beginning. This scenario would be possible in a variety of situations; it is possible to imagine in the case of a performer who begins with nervousness. Excerpt 2 was designed to resemble a generally good performance marred by a few mistakes at the end; possible in a variety of "real" situations. This scenario is most likely in the case of a performer who "lets down" performance attentiveness before the end of the piece. Excerpt 3 was designed to resemble a performance wherein the performer appears to be much more comfortable with Phrase A than with Phrase B. As in most performances, in this scenario there is a learning effect that allows for improvement during the repetition of Phrase B. Excerpt 4 represents a scenario where the performance worsens as it progresses. This kind of situation is perhaps, most noticeable in performers who are thrown by mistakes, allowing each mistake to compound and further distract from the remainder of the performance. It is also feasible to imagine this scenario occurring in a performer who "gets out of his/her head" and becomes increasingly more nervous. Excerpt 5 represents a scenario that is the antithesis of excerpt 4. In this case, the performance improves as the performer plays, perhaps because the act of playing itself relaxes the performer.

These five excerpts were evaluated by a panel of experts who had a total of 59 years of performing or conducting experience. Each expert listened to the recording and then paired the description of the excerpt found above with the excerpt itself. Experts were allowed to listen to the recording as many times as needed, although typically the task was accomplished after only 4 repetitions. Experts were able to match descriptions with performances with 100% accuracy.

Design

Each excerpt represented a different experimental condition, and subjects heard all five excerpts (within-subjects design). No practice example was given, because it was assumed that presenting the "ideal" model prior to the administration of the experimental conditions would have incontrovertibly subsequent listening. Given this limitation, some manner of order effect was possible, as subjects would (to some degree) compare each excerpt to the one heard first.

To attempt to address order effect, four randomized orders were created of the five experimental conditions. Subjects heard the excerpts in one of the four orders. Group 1 (n=30) heard excerpts in order 2,5,4,3,1; group 2 (n=32) heard excerpts in order 4,2,3,1,5; group 3 (n=36) heard excerpts in order 1,2,3,5,4; and group 4 (n=31) heard excerpts in order 4,2,1,3,5. The orders were created using a web-based research randomizer application (Urbaniak & Pious, 2009).

In addition, a short unrelated Renaissance musical excerpt (Attaignant, 2006) was placed between each condition. This music was 15 seconds in length, and was separated from the experimental conditions by approximately three seconds of silence. This followed a protocol established in prior research (Duke, Geringer, & Madsen, 1988) in that it ". . . was intended to inhibit immediate and direct comparison of pitch . . . in subsequent examples" (p. 111). The total time required for each of the five conditions, plus the intercessory material, was just under seven minutes.

Four separate CDs were created, each of which contained one of the randomized orders. The stimuli were played on an RCA DRC-220 DVD/CD player. Subjects heard the stimuli at individual listening stations, each of which was partitioned from the view of other subjects and equipped with a pair of AKG K-2405 Semi-Open Studio Headphones. A copy of the instructions and a CRDI dial was at each listening station. The dial was configured so that when it pointed straight up it was in a neutral position. This corresponded with unit 128 on the device. The left hand side of the dial was labeled "negative" and encompassed units 0-127, while the right hand side of the dial was labeled "positive" and encompassed units 129-256.

Subjects were admitted to the testing area, and asked to complete a consent form and a demographic form. Subjects were then instructed to follow along as the instructions for the study were read:

> In a moment, you will hear five different performances of the same composition. It will be played on solo cello, and will last for 1 minute and 22 seconds. As you listen to each performance, please turn the dial in front of you to indicate your perception of the QUALITY OF THAT PEFORMANCE. Before beginning each listening task, make sure your

> dial is in the neutral position. There will be a short break (14 seconds) between each performance during which music unrelated to the study will play. Use this time to return your dial to the neutral position. Please make sure your dial is in the neutral position before we begin.

Subjects manipulated the CRDI dial while listening to the stimulus CD. At the conclusion of the listening task, subjects were instructed that they could complete the free response section of the demographic form or exit.

CHAPTER 4
RESULTS

Data were gathered in numerical format via the CRDI, which time-sampled the responses of the participants every ½ second. This resulted in just over 160 ratings for each excerpt by each subject, or a total of just under 21,500 data points for each excerpt. Mean ratings and standard deviations for all subjects and for each excerpt were charted graphically (composite). Further, mean ratings and standard deviations for all subjects and for each excerpt were charted within each order (order-specific).

Visual analysis of the composite mean ratings and standard deviations indicated a strong relationship between the excerpts' responses and the musical content of Excerpts 1, 2, 3, and 5. Excerpt 4, when examined from a composite perspective, appeared to have not functioned as intended. The consistently rising standard deviation of Excerpt 4, however, seemed to indicate that further analysis was prudent.

Visual analysis of the order-specific mean ratings and standard deviations confirmed the results of the composite analysis for excerpts 1, 2, 3, and 5. It also suggested a rationale that helped to explain the aberrational data for excerpt 4. Excerpt 4 appeared in the first position in two of the randomly generated orders (orders 2 & 4), and when these data were examined in isolation, it was apparent that Excerpt 4 was influenced by a strong order effect. Specifically, when Excerpt 4 appeared in the first position its ratings were decidedly higher than when Excerpt 4 was found in any other position.

While visual analysis of the graphic material makes it appear that subjects used a very limited portion of the available dial, there were subjects that used the entire dial (from 0 to 255). In fact, examination of individual subject data suggests noticeable differences in the way that subjects approached the task.

Graphic Analysis of Composite Ratings

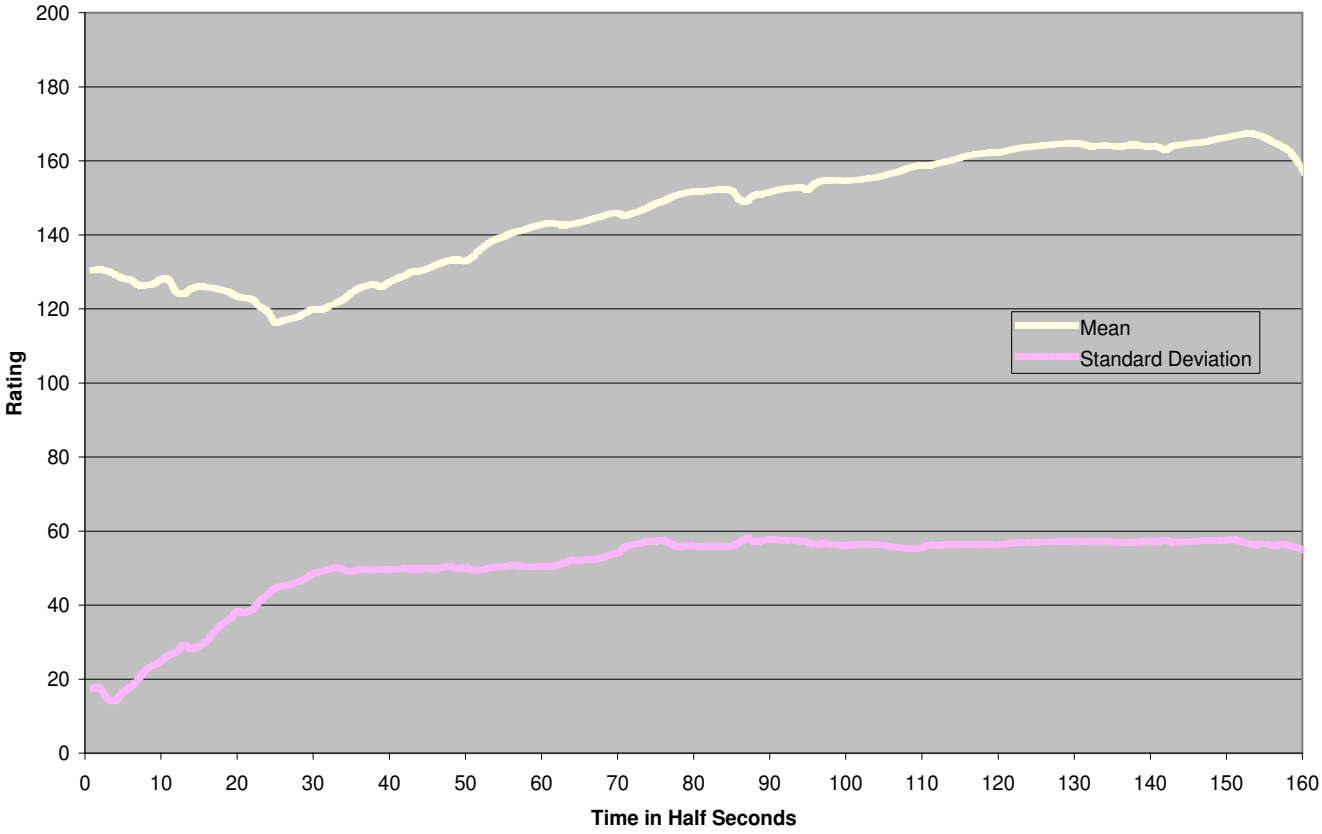

Figure 3. Graph of Composite Means and Standard Deviations for Excerpt 1 across time (low-error, good, good, good).

The graph of the composite means and standard deviations for Excerpt 1 (low-error, good, good, good) seems to suggest a response to the "low-error" segment of the composition. The segment contains two distinct errors, one which occurs at the 3 second mark (6 above), and another which occurs at the 9 second mark (18 above). Based on a visual examination of the graph above, it would appear that after these two errors occur, the mean subjects' ratings do not return to the neutral position until approximately 22 seconds into the excerpt, which is roughly the last measure of the A section, 2nd time. This is salient because only the first segment of excerpt 1 was altered; from the 12 second

mark forward the performance is "good." The minimum mean score for excerpt one was 116.40; the maximum was 167.40 (a range of 51 points).

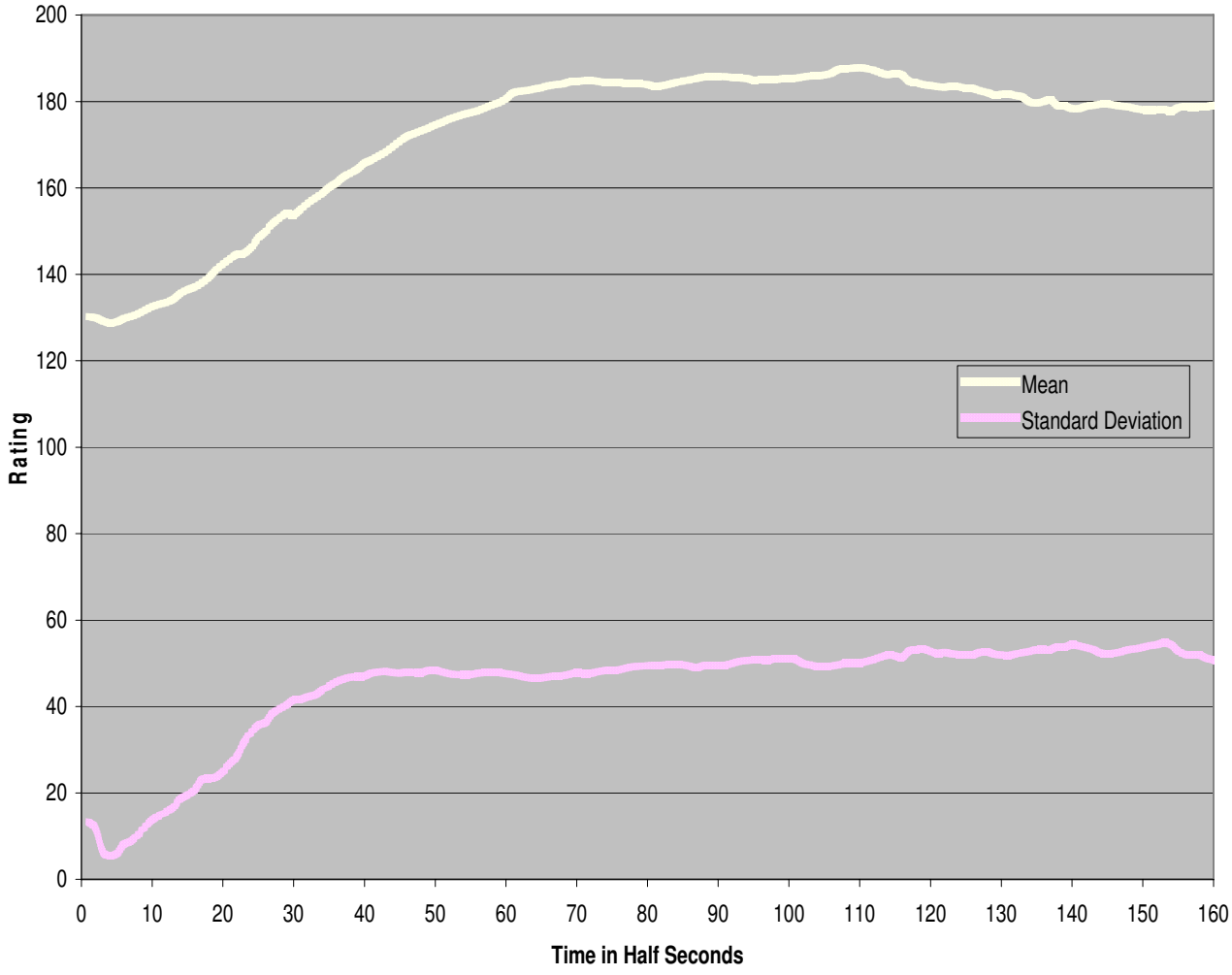

Figure 4. Graph of Composite Means and Standard Deviations for Excerpt 2 across time (good, good, good, low-error).

The composite graph of Excerpt 2 (good, good, good, low-error) seems to suggest a steady increase in preference for the composition, with a slight decline occurring during the "low-error" segment in the repeat of phrase B. The errors in Excerpt 2 occur at the 57 second mark (114 half-seconds) and the 1 minute, 8 second mark (136 half-seconds). Visual evaluation of the graph shows that the decline is very slight, accounting for a

range of less than 10 units. The minimum mean score for excerpt two was 128.70; the maximum was 187.84 (range of 59.12 points).

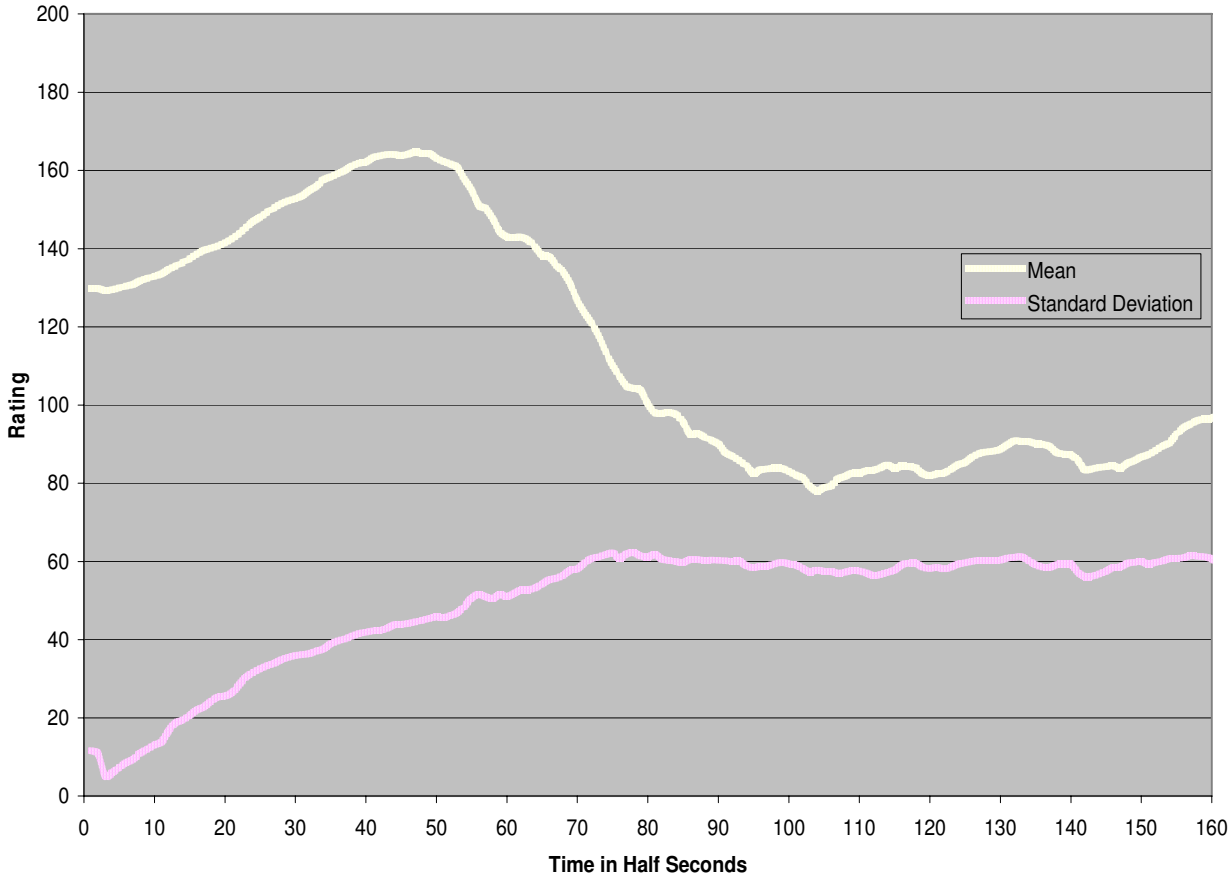

Figure 5. Graph of Composite Means and Standard Deviations for Excerpt 3 across time (good, good, high-error, low-error).

The graph of the composite ratings for Excerpt 3 (good, good, high-error, low-error) seems to indicate a gradual increase in preference until approximately 24 seconds (50 half-seconds), at which point the ratings drop severely. Beginning at the 24 second mark, errors occur almost every three seconds until the 49 second mark (98 half-seconds). From the 98 half-second mark forward, there are only 2 errors in the remainder of the excerpt. Despite this relative lack of error, ratings show little sign of recovery, although the graph could be interpreted to indicate a very slight recovery towards the

very end of the excerpt. The minimum mean score for excerpt three was 78.06; the maximum was 164.67 (a range of 86.60). This minimum mean score was the lowest of any excerpt in the study.

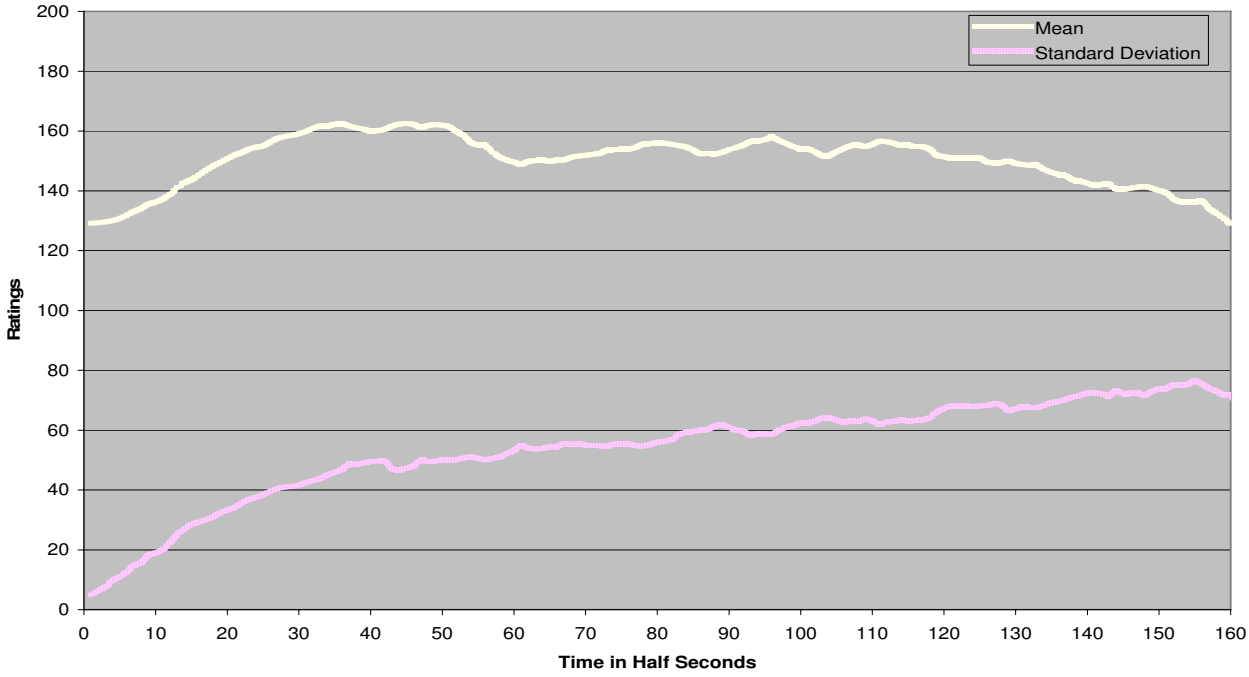

Figure 6. Graph of Composite Means and Standard Deviations for Excerpt 4 across time (good, low-error, low-error, high-error).

The graph of the composite ratings of Excerpt 4 (good, low-error, low-error, high-error), upon visual examination, seems to respond in a fashion dissimilar to any of the other excerpts. The first error in Excerpt 4 occurs at the 18 second mark (36 above) while the second error occurs at the 21 second mark (42 above). A slight decline is apparent after these errors, coinciding with the first of the three errors that occur during the low-error section of phrase B at 0.27 (54 above). After this slight decline, however, the subjects' ratings show some signs of mild recovery, despite further errors at 0.40 (80 above), and 0.49 (98 above). The high error segment begins at 0.52 (124 above). While the excerpt shows a downward trend for the rest of its duration, the decline is not as severe as the one created by the same segment in Excerpt 3. This excerpt was also the only one of the five to display a rising standard deviation. The minimum mean score for

excerpt four was 129.15; the maximum was 162.47 (a range of 33.33). This was the smallest range of any of the composite means.

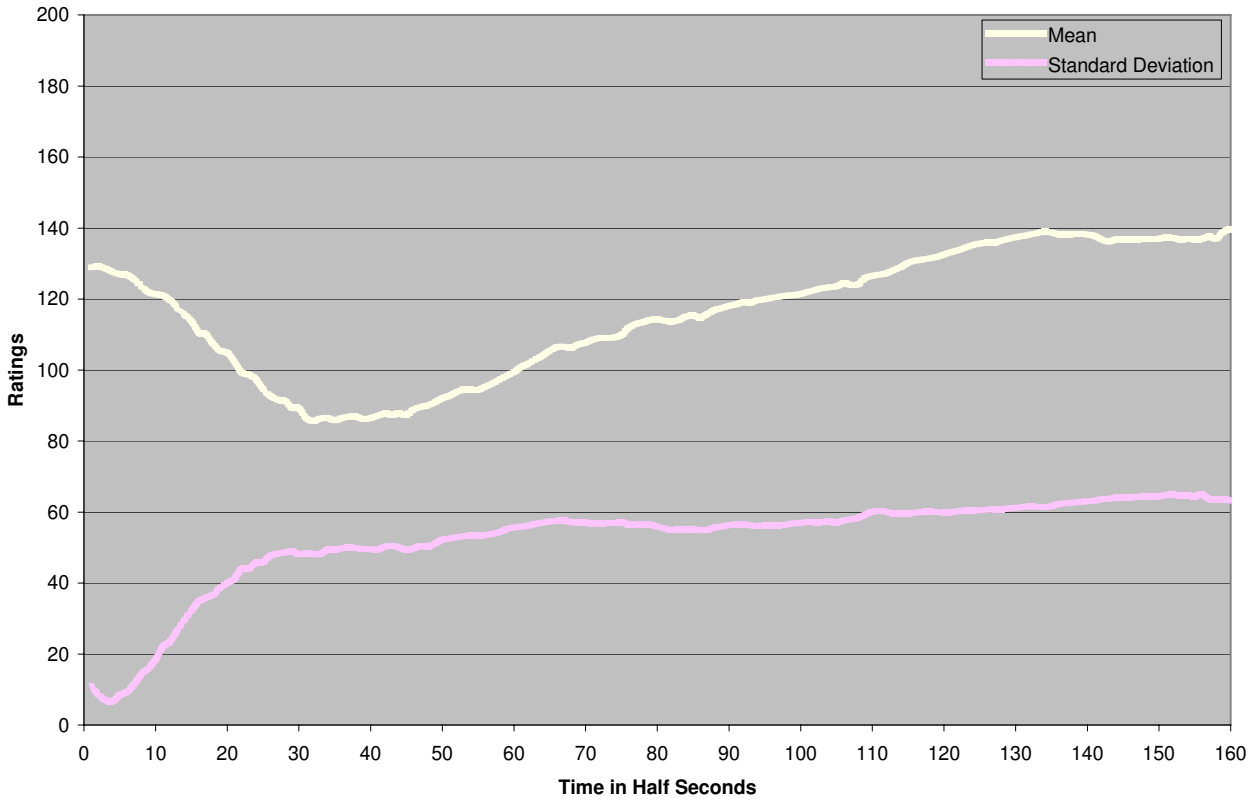

Figure 7. Graph of Composite Means and Standard Deviations for Excerpt 5 across time (high-error, low-error, good, good).

The graph of the composite mean and standard deviation for Excerpt 5 (high-error, low-error, good, good) displays a response to the frequent errors of the first segment followed by a long, gradual recovery. The ratings seem to suggest an immediate negative reaction to the high-error segment of the excerpt. This negative reaction causes a decline which persists until 0.16 (32 above), midway through the low-error segment. At this point the ratings show a steady upward trend. From approximately 0.22 (44 above) forward the performance is unaltered, yet the ratings do not return to neutral until almost 0.58 (116 above), or midway through the repeat of the B phrase. The minimum mean

rating for Excerpt 5 was 85.65; the maximum was 139.70 (a range of 54.05). This maximum mean rating is the lowest of any excerpt.

Graphic Analysis of Order-Specific Ratings

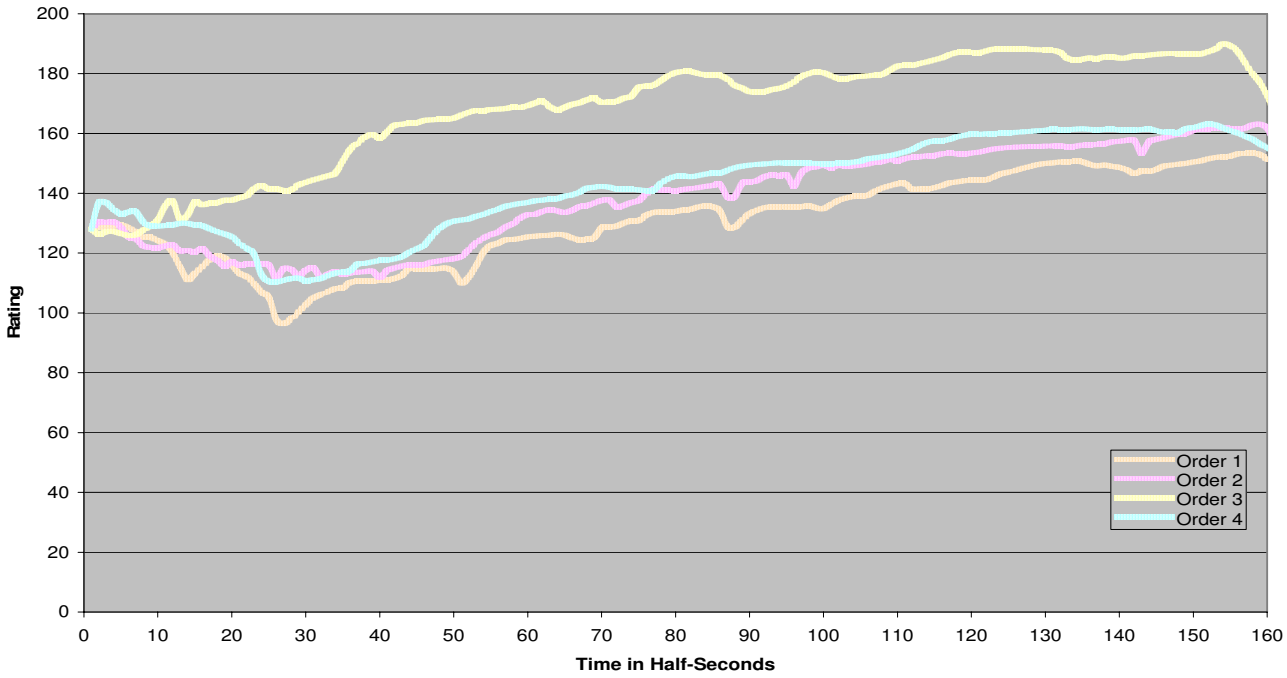

Figure 8. Graph of Order-Specific Means for Excerpt 1 (low-error, good, good, good) across time.

Visual examination of the graph of the order-specific ratings for Excerpt 1 reveals that Excerpt 1 was rated uniquely in Order 3. Order 3 presented Excerpt 1 in the first position, which functioned to inflate its overall ratings. When Excerpt 1 was heard in the first position, its mean, maximum rating was 189.91; its minimum was 125.86. This means that its mean rating only dipped slightly below neutral (128) and that it had a range of 64.06. By contrast, when Excerpt 1 was heard in position 3, 4, or 5 (as in orders 1, 2, and 4) it had a mean, maximum rating of 163.19 and a minimum rating of 93.37. The range of Excerpt 1 in the other three orders was 66.83, which is somewhat similar to the range in Order 3 (64.06), but visual examination of the graph seems to suggest that

subjects who heard the excerpt in the first position did not respond to the low-error segment at the onset of the excerpt.

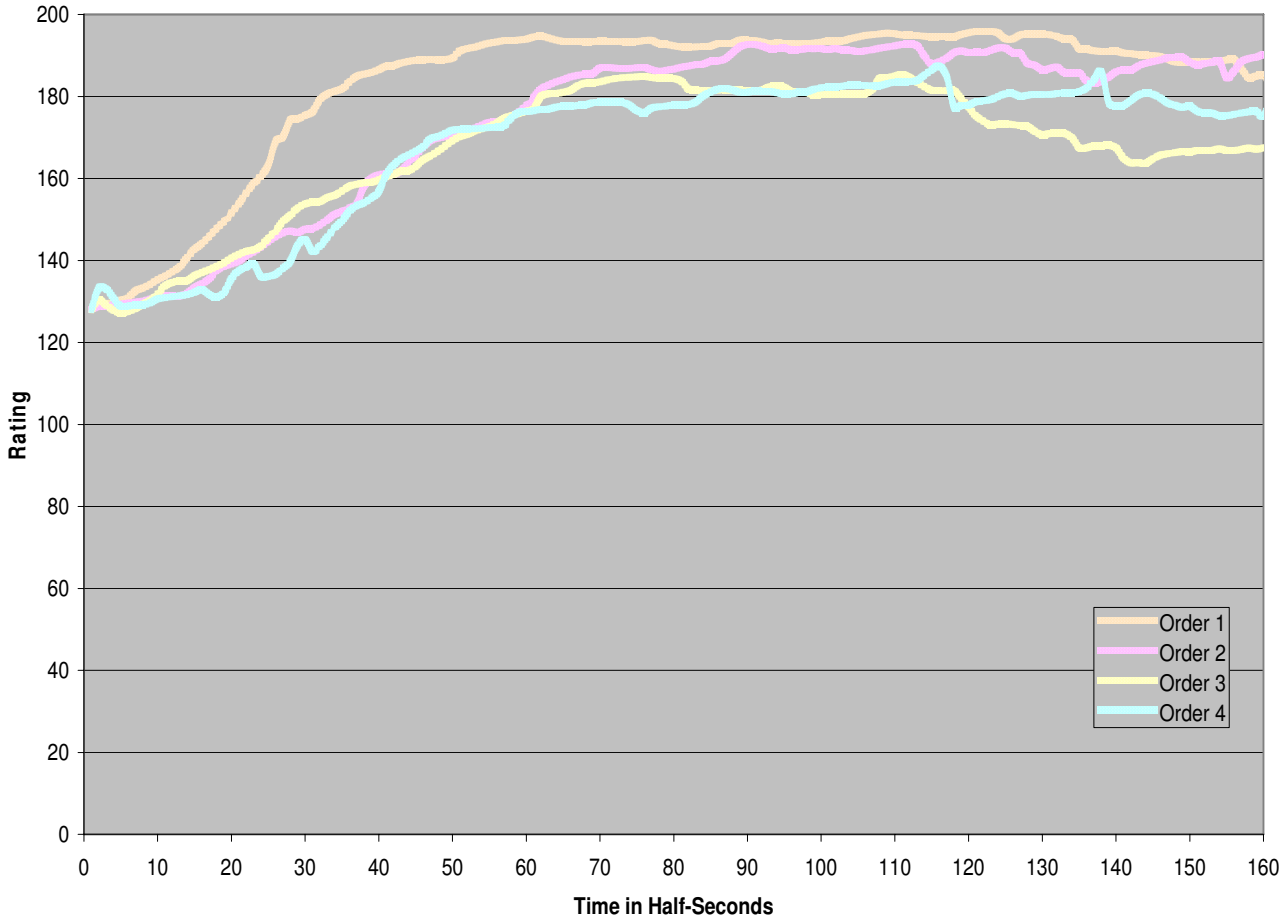

Figure 9. Graph of Order-Specific Means for Excerpt 2 (good, good, good, low-error) across time.

Visual examination of *figure 9* seems to suggest that Excerpt 2 received slightly more positive ratings at the onset of the excerpt in Order 1, when it was in the first position, than in any other order. Subjects who heard Excerpt 2 in the 1st position reached a mean rating of 180 by approximately 0.18 (36 half-seconds), which corresponds with the 1st ending of phrase A. Subjects who heard Excerpt 2 in the 2nd position, on the other hand (where it was in all other orders), did not reach a mean rating of 180 until approximately 0.33 (66 half-seconds). Despite this similarity, Excerpt 2 showed more

variability in Order 3 than in Orders 2 or 4. In Order 3, however, excerpt 2 followed Excerpt 1. In Orders 2 and 4, it followed Excerpt 4.

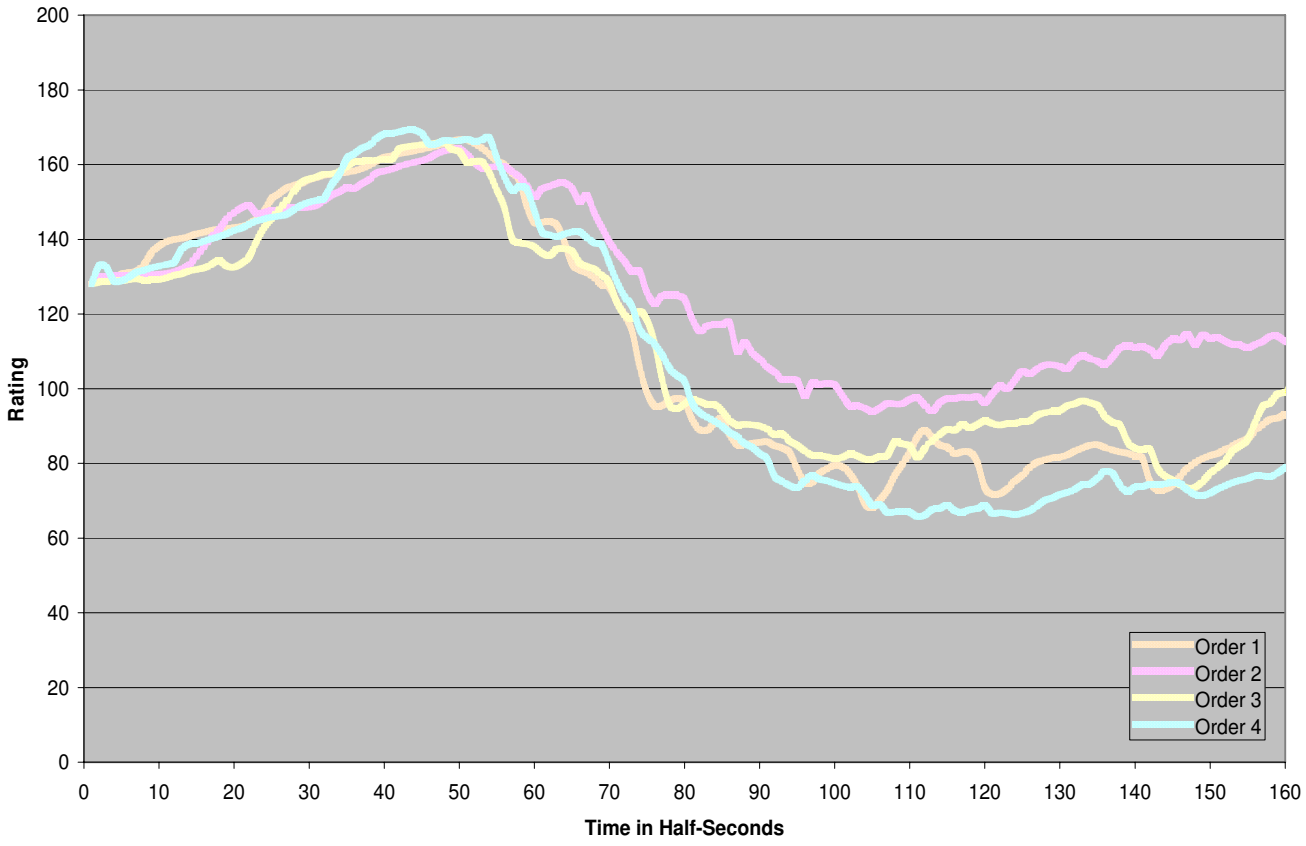

Figure 10. Graph of Order-Specific Means for Excerpt 3 (good, good, high-error, low-error) across time.

The order-specific graph of Excerpt 3 (Figure 10) generally shows a similar shape across all four orders. Excerpt 3 was in positions 4, 3, 3, and 4. The range between Orders 4 and 2, particularly towards the end of the excerpt, could be attributable to an order-interaction. In Order 2, Excerpt 3 follows Excerpt 2. Subjects who heard the excerpts in this order would not yet have been presented with a completely error-free B section. In Order 4, Excerpt 3 follows Excerpt 1. Subjects who experienced the excerpts in this order would have heard an exemplary second section from which to make comparisons. This discrepancy appears to display itself in the mean minimum score and range of the ratings.

The minimum score of Orders 1, 3, and 4 was 65.81; whereas the minimum score of Order 2 was 93.91. The range of Orders 1, 3, and 4 was 103.48; whereas the range of Order 2 was 70.38.

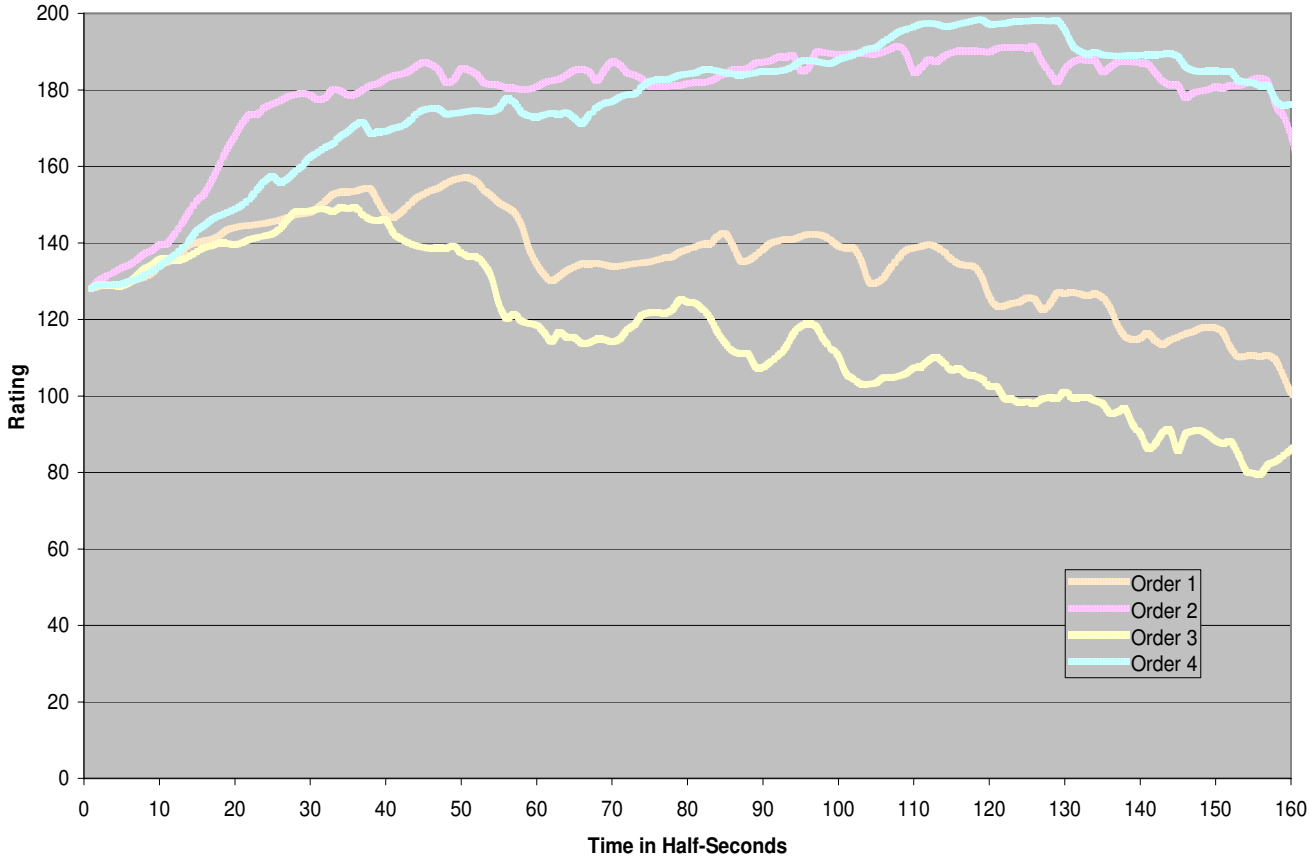

Figure 11. Graph of Order-Specific Means for Excerpt 4 (good, low-error, low-error, high-error) across time.

Figure 11, upon visual examination, seems to exhibit the most obvious signs of an order-interaction. Excerpt 4 was in position 1 for two of the experimental conditions (Orders 2 & 4) and in this position it received much more positive ratings. In Orders 1 and 3, Excerpt 4 was in positions 3 and 5 (respectively) and was judged much more harshly. Subjects who heard Excerpt 4 in position 5, having heard all of the previous experimental conditions, judged Excerpt 4 most negatively. In Orders 2 and 4 (position 1) Excerpt 2 had a maximum mean rating of 198.32, a minimum of 129.06, and a range of

69.26. In Orders 1 and 3 (positions 3 & 5, respectively) Excerpt 2 had a maximum mean rating of 156.97, a minimum of 79.66, and a range of 77.31.

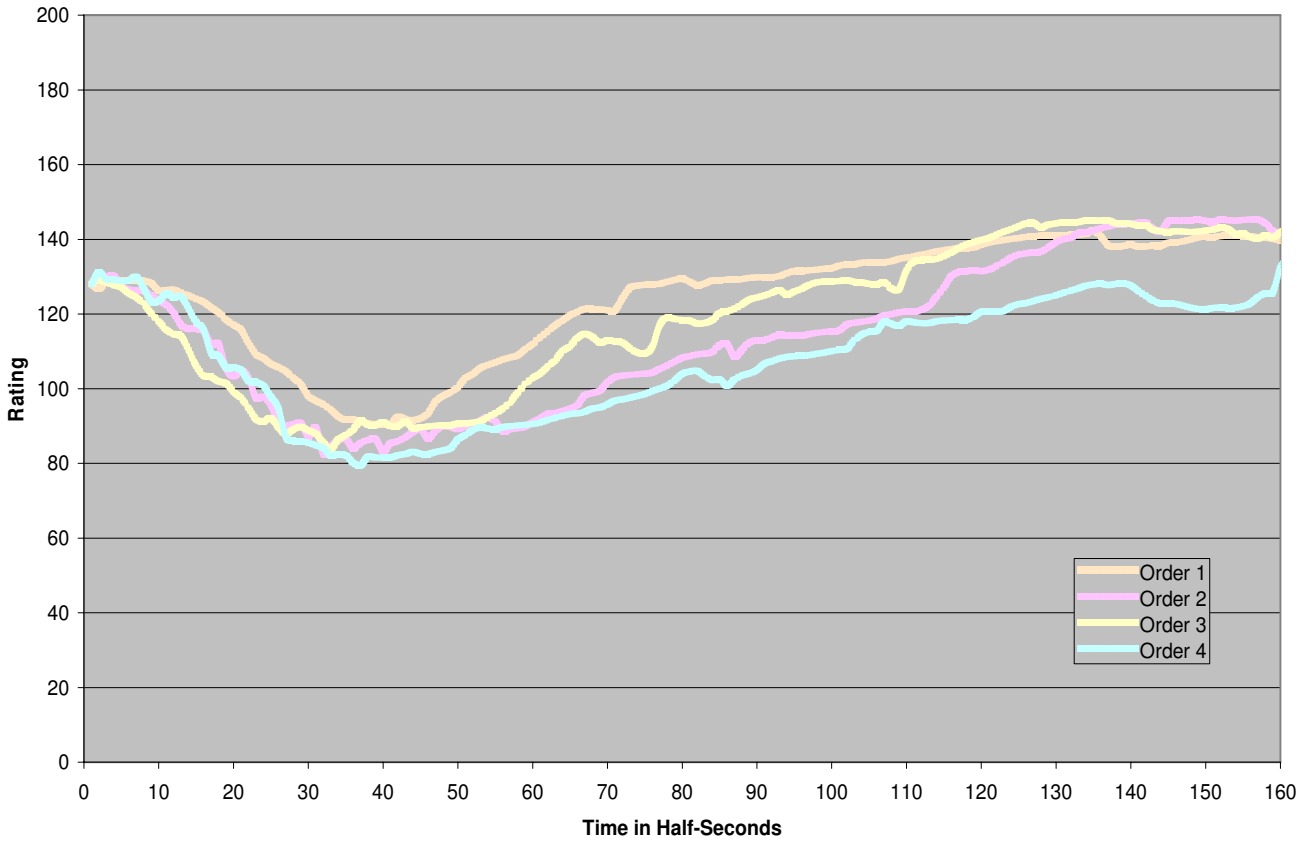

Figure 12. Graph of Order-Specific Means for Excerpt 5 (high-error, low-error, good, good) across time.

The order-specific Excerpt 5 graph (Figure 12) shows a similar ratings trend for Excerpt 5 in every position. Excerpt 5 appeared in positions 2, 5, 4, and 5. It appeared to fare better in position 2 (Order 1 above) than in any other position, although this improvement seems to translate as quicker recovery time. When Excerpt 5 was heard in position 2, it returned to neutral (128) by approximately 0.40 seconds, approximately 20 seconds prior to the mean recovery to neutral when it was heard in positions 4 or 5.

CHAPTER 5
DISCUSSION

Examination of the data reveals several salient factors that appear to have affected the study. Of these, the two most obvious are the presence of an order-effect and a learning factor (short-term maturation). The learning factor appeared to function such that subjects became more discriminating as the task progressed. Given this reality, excerpts heard first were more likely to garner positive responses, while excerpts heard later were more likely to elicit negative reactions. The strongest example of this occurred in Excerpt 4, which appeared in position 1 twice, but this effect was obvious for every excerpt in position 1. Even with these factors addressed, it seems clear that the subjects' perception of the excerpts across time did display responsiveness to error. It is also apparent that, while subjects' perceptions did recover from error, the time required to recover was long given the brevity of the mistake(s) and its/their context in the excerpt.

A distinctive example of this phenomenon occurs in the graph of Excerpt 1 (Figure 13). As was previously mentioned, subjects mean ratings of Excerpt 1 dipped to their lowest point at approximately 12 seconds into the excerpt. From this point, they showed steady signs of increase but did not rise above neutral for another 20 seconds. This long incline is in contrast to the decline created by response to the error, which took almost half that time (approximately 11 seconds). The strength of negative stimuli as compared to positive stimuli could offer one explanation for this phenomenon. If negative stimuli possess more magnitude than positive stimuli, it is reasonable to assume that they would impact assessment more consequentially.

Accounting for the learning factor, this magnitude becomes more apparent. Excerpt 1 appeared in the 1st position in Order 3, and when these mean scores are removed from the total the response to Excerpt 1 is more negative.

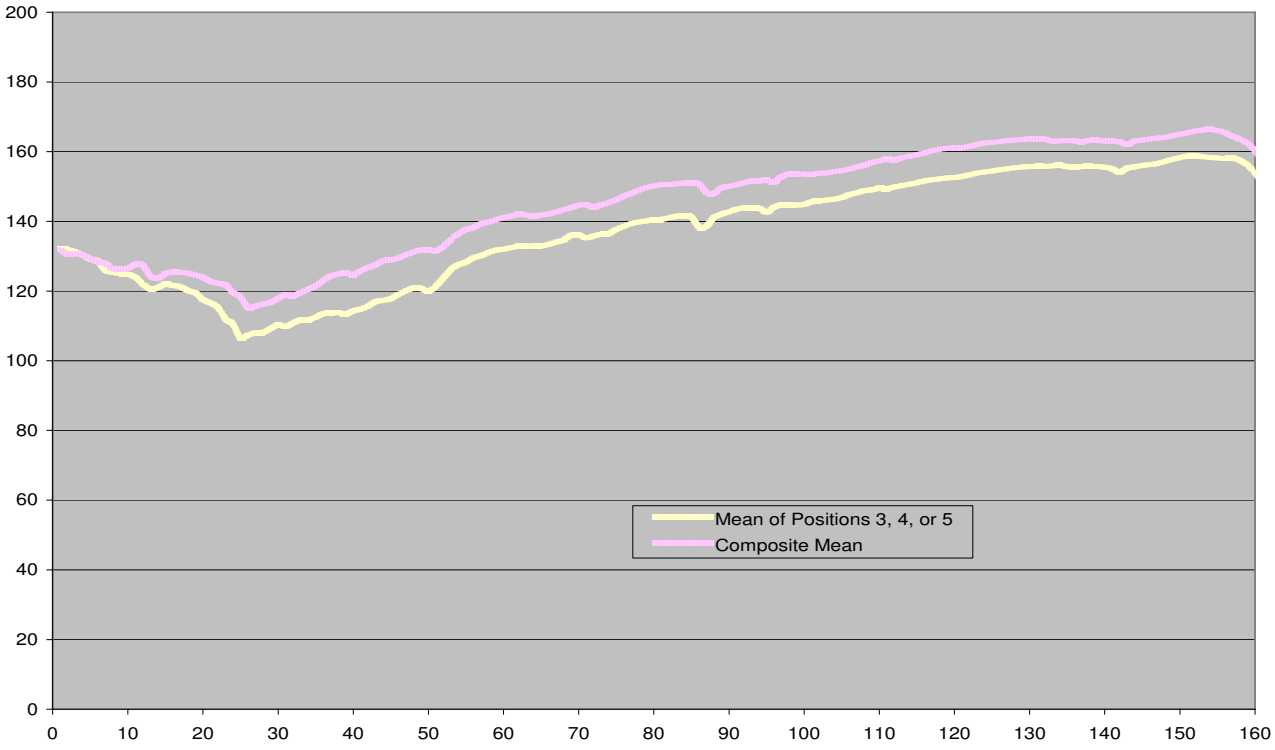

Figure 13. Graphic Analysis of Excerpt 1, with Composite Means and Means of Excerpt when not in position 1.

Figure 13 visually represents this phenomenon. The composite mean, which includes the ratings of Excerpt 1 in 1st position, is obviously higher than the mean of Excerpt 1 when it is in positions 3, 4, or 5. As was previously noted, subjects who heard Excerpt 1 in the first position were less affected by the errors at the beginning of the excerpt. This inflated their overall perception of the excerpt, and in turn, the composite mean ratings. When this is taken into account, it is clear that the decline caused by reaction to the error at the beginning of the excerpt is steeper (a difference of roughly 9.5 points) and that the peak mean rating suffers as well (it is roughly 4 points lower). More compelling, however, is that it is apparent that the recovery took much more time (a difference of almost 15 seconds).

Viewed in light of these adjustments, it could be inferred that in the case of Excerpt 1, subjects reacted quickly and decisively to error, while error-recovery was more gradual. This would seem to suggest that errors have residual effects that persist after they have been "passed" chronologically. There is also some evidence to suggest

that negative performance characteristics possess more strength than positive performance characteristics.

These tendencies are only somewhat apparent in the case of an excerpt that was predominantly well-performed. Excerpt 2 featured only slight mistakes situated at the conclusion of the excerpt. As such, subjects heard almost a full minute of "good" performance before encountering error (in an excerpt that was 1.22 long). Despite this positive proclivity, subjects still reacted strongly to negative stimuli (if not as strongly as in Excerpt 1). Excerpt 2 was heard in the first position in one order, but was heard in the second position in all other orders. When the means of Excerpt 2 are graphed without the first position example included, there are only slight differences in the ratings.

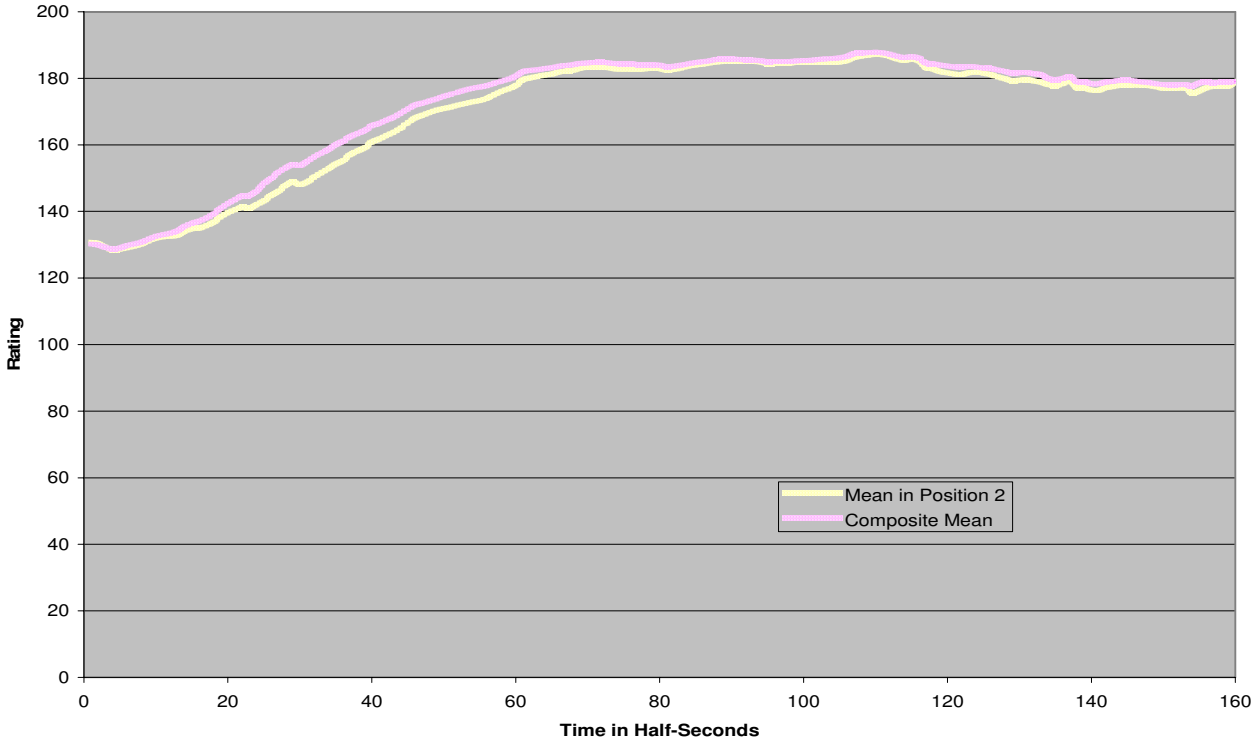

Figure 14. Graphic Analysis of Excerpt 2, with Composite Means and Means of Excerpt when not in position 1.

This graph indicates that listeners who heard Excerpt 2 in position 2 were more critical of the performance. It is interesting to note that while the errors at the conclusion of the excerpt begin a downward trend, it is not nearly as precipitous as the decline

caused by the errors at the beginning of Excerpt 2. Several questions arise when these results are considered in tandem with Excerpt 1: Do errors at the beginning of an excerpt have more strength than errors at the conclusion of an excerpt? And, would ratings taken moments after the conclusion of Excerpt 2 continue to evidence a downward trend? Or, can the final ratings of Excerpt 2 be considered a plateau? Research in serial position would seem to suggest that errors at the beginning or end of the excerpt would be equally important. That does not appear to be the case, however, as far as these two excerpts are concerned.

Excerpt 1 was rated, at its peak, slightly above 160; whereas Excerpt 2, at its peak, was rated slightly below 190. This is especially curious when one considers the amount of similar material heard by the subjects. The area from 0.10 to 0.56 was completely unaltered in either excerpt. This means that the subjects heard an identical recording for roughly 56% of the excerpt. Of course, the errors in Excerpt 1 do occur closer together, given the shorter length of Phrase A. This could account for some of the variance in the scoring, but it is possible that a "first impression" effect could be subtly influencing outcomes as well.

Examination of the data attached to Excerpt 3 lends credence to several of the rationales proposed thus far. First, the negative segments of Excerpt 3 produced ratings of considerable strength. These ratings were the lowest composite means of any in the study and effectively negated the gains of the prior sections. Part of the explanation for this may be found in the contrast presented by Excerpt 3. Listeners heard 0.24 of unaltered performance before errors began, but then they heard an error almost every three seconds. This level of contrast in performance is probably somewhat rare, and listeners responded immediately and emphatically to the change. Even this argument does not appear to account for the severity of the change, however; when the discrepancy in ratings is thoroughly examined. Participants' ratings during the good portion of Excerpt 3 increased gradually by just over 30 points. Ratings during the high-error section, by contrast, dropped nearly 80 points. Such a paradox seems to lend credibility to the notion that negatives in performance influence listeners with much more strength than do positives.

Second, the negative ratings of Excerpt 3 were very slow to recover. The recovery was so slow that it could even be argued that no recovery took place at all. This is especially pertinent when the last segment of Excerpt 3 is taken into account. For approximately the last 30 seconds of Excerpt 3, there are only 3 alterations in the performance. Given that information, it would seem reasonable to assume that ratings would improve. The improvement that occurs is, however, incredibly slight. A similar segment at the onset of Excerpt 1, however, took upwards of 15 seconds to improve after subjects heard its final mistake. This seems to suggest that there was not enough time at the conclusion of this excerpt to allow for the sort of gradual recovery that error seems to require.

Excerpt 4 was the most problematic of the experimental conditions. The order-specific graph seems to suggest that Excerpt 4 was subjected to a strong order-effect. Comparing the composite mean of Excerpt 4 with the mean of Excerpt 4 when it was not in first position yields interesting results.

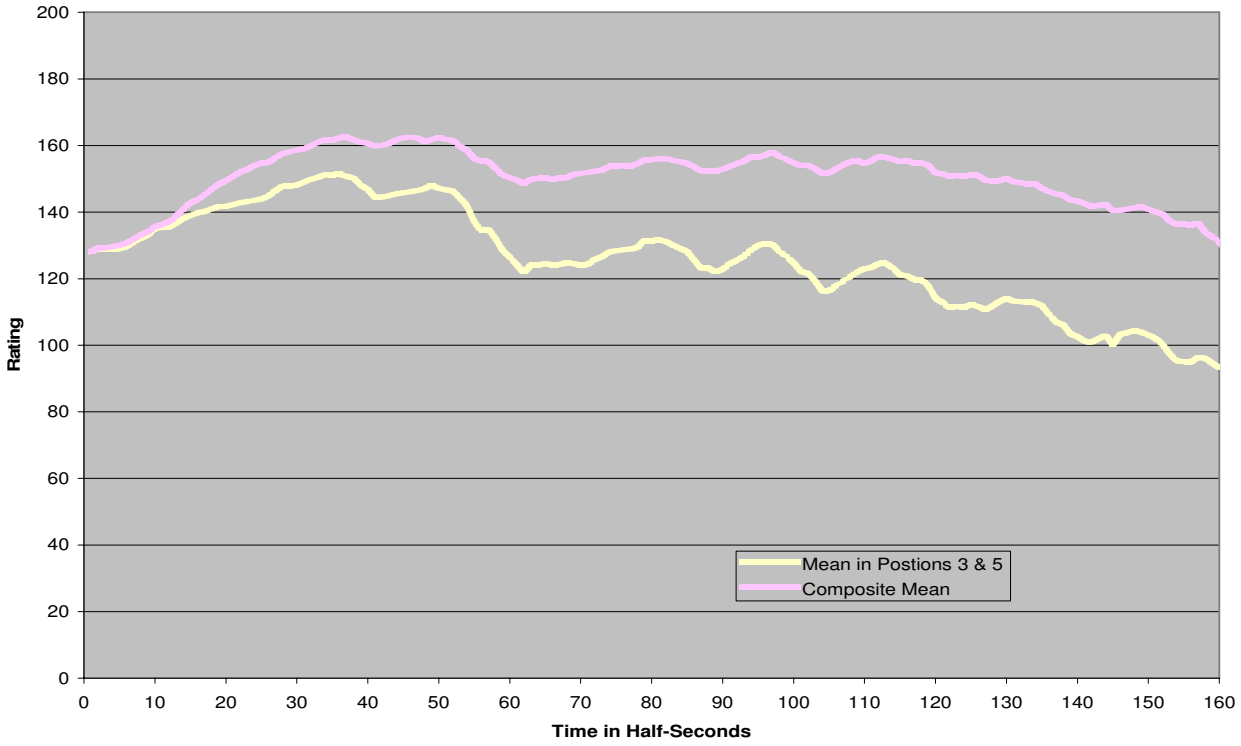

Figure 15. Graphic Analysis of Excerpt 4, with Composite Means and Means of Excerpt when not in position 1.

If Excerpt 4 is graphed excluding the data gathered when it was in first position, it becomes clear that the subjects' overall response to this Excerpt was increasingly negative. Subjects responded to the unaltered segment of the excerpt with a slight increase in rating, and then responded sharply when the low-error segment began. As with Excerpt 3, the disparity between good performance and bad performance seemed to provoke a very strong reaction. In contrast to Excerpt 3, the negative trend of Excerpt 4 was not judged as harshly as the negative segments of Excerpt 3. Strict comparison may not be entirely reasonable, however, because Excerpt 3 presents a high-error segment in the middle of the piece, while Excerpt 4 presents a high-error section at the conclusion of the piece. Given the relative latency of response to error demonstrated thus far, it may be that a measure taken after the conclusion of Excerpt 4 would have shown a continued pattern of decline.

Thus far, the findings seem to suggest that positive and negative aspects of performance have different latency effects. Subjects seem to respond slower to positive stimuli (a longer period of latency), while negative stimuli provoke more immediate response. Additionally, the latency of negative stimuli seem to be much stronger than that of positive stimuli. Excerpt 5 demonstrates these findings in a salient fashion.

Excerpt 5 was intended to simulate a performer who had a "bad-start," but gradually recovered over time. Subjects uniformly responded to the Excerpt, although subjects who heard the Excerpt in position 2 were more positive overall. In some ways, Excerpt 5 seems to be the most suitable example of the positive/negative latency effect documented by this study. Subjects responded to the negative stimuli at the beginning of the Excerpt swiftly, and then gradually recovered as the Excerpt became more positive. A strong case for the strength of the negative can also be made by the response to Excerpt 5. Subjects heard 22 seconds of altered performance, followed by 1 minute of good performance. Despite this preponderance of good, the peak rating of Excerpt 5 was the lowest of any experimental condition. The response to Excerpt 5 would also seem to indicate that subjects tend to view performances with poor beginnings negatively, although not necessarily worse than when errors are encountered in other contexts. This

seems to suggest that a primacy effect could be occurring, but it is unwise to draw any definitive conclusions on the strength of one study.

Practical Applications

Several cogent points emerge as these data are put in context. First, from a performance viewpoint, it is clear that trained listeners do possess the capacity to identify and respond to errors as they occur. This is not new information in the practical sense; most trained musicians would acknowledge that this is part of their experience. There is also precedent for this finding in error-detection research (Brand & Burnsed, 1981; Deal, 1985; Killian, 1991; Larson, 1977; Sheldon, 2004), and this study (while not error-detection research *per se*) could be seen as another empirical verification that musicians are capable of identifying musical errors in simple contexts. These findings suggest that performers would be well-advised to prepare music such that minimal errors occur in performance, because trained listeners will identify and respond to errors.

Second, negative musical stimuli seem to possess more magnitude than positive musical stimuli. Outside of musical contexts, this trait is not uncommon. It is the nature of the human experience to attempt to codify and categorize our environments, such that it is easier to "make sense" of our world. When undertaking this process, our attention is naturally drawn to that which is different, or wrong. The wrong stands out (possesses more magnitude) because it is different. Anyone who has ever had a single scene 'ruin a movie', or had a single experience 'ruin a date,' can attest to the strength of the negative. Within behavioral psychology, there are ratios that attempt to explain this phenomenon, and many people are familiar with the 5:1 ratio of positive comments to negative comments advised by counselors, etc.

It is somewhat unusual, however, to document a similar effect in music. Many musicians would probably acknowledge that negative musical stimuli possess strength, as they would with error detection, but the magnitude suggested by this investigation may be beyond that which most musicians would expect. The magnitude of negative stimulus in this study (particularly in high-error examples) was strong enough that it frequently cancelled out or greatly hindered the effects of all subsequent, positive stimuli. Excerpt 5, for example, consisted of 27 % negative stimulus from a time perspective. While this is not a 5:1 ratio, it does represent a predominantly positive listening experience (73%), yet

subjects barely rated this performance higher than neutral by its conclusion. The implications for these findings are obvious: musicians should attempt to minimize error, because even the smallest errors appear to possess more strength than positive stimuli of similar duration.

Regarding latency, it is clear that musicians should possess an understanding of just how much time is required by listeners to mentally recover from an error. These findings suggest that recovery time is considerable, and often occurs long after the error has passed. In some ways, these findings may seem to be in contradiction to those reported by Southall (2003). He found that while his distractors did function as intended, subjects ". . . evidenced a quick recovery and continued to have an aesthetic response following the periods of distraction" (p. ix). First, it should be noted that Southall's investigation and the current study have a notable difference: Southall inserted non-musical elements into a musical excerpt, whereas the current study attempted to fabricate a musical performance by inserting wrong notes into a musical excerpt. Findings of the two studies must be considered in light of this reality. Second, the recovery documented by Southall was quick within the context of a specific musical performance, in this case, an excerpt from *La Boheme* that was over 10 minutes in duration. And while that recovery time was quick, it still required approximately 50 seconds per distractor.

Several venues for future investigation are available. While the present study did not present a practice example for aforementioned reasons, it is not clear to what extent familiarity could have influenced these findings. Practice examples by other performers (not Yo-Yo Ma) could be played prior to the administration of the experimental conditions for the purposes of familiarizing all subjects with the composition. This may still produce bias, however, as there are sure to be differences in interpretation present. Presenting subjects with a visual representation of the score could be implemented as well. This may sensitize those less familiar with the recording to the performance, but it may also have the negative effect of changing the nature of the experiment. Part of the challenge of this study was structuring the instructions and examples carefully such that the experiment did not become one large error-detection exercise. Providing subjects with the score could substantially move the study in that direction.

Still another question posed by the study concerns the nature of the errors utilized. Not all errors are created equal, and while the examples used in this study were evaluated by experts, no attempt was made to address the magnitude of certain errors compared with others. Important questions raised by this investigation are: what sorts of errors provoke the most negative response? Do different listeners perceive errors within differing hierarchies? The current study made use only of errors in pitch. Would errors of rhythm or tone provoke similar responses? The technology available to music researchers has reached a point where answers to these sorts of questions may soon be forthcoming.

Sample differences could also bear investigation. This study was primarily concerned with the reaction of trained musicians to errors in time, but no attempt was made to examine how specific subjects responded to the experimental conditions. Would string players react more harshly to wrong notes of wind players? How do players of unfretted instruments (such as violin, viola, cello, bass, trombone, and one could argue, the voice) differ in their evaluation of examples of this kind from other musicians? Does age matter? Or music specialty? It could be possible that performers would view errors in time differently than educators.

There is also the question of literature to consider. The present study used one very specific recording to create all stimulus examples. Therefore, the generalizability of the study is somewhat limited. Further application of this approach, to other styles, genres, instrument families, and time periods will undoubtedly reveal more about this process.

The latency, or recovery, effect demonstrated in this investigation probably would be the most pertinent issue for examination in future research. It would be worthwhile to replicate the current study with the addition of a post-hoc, static measure after each excerpt has finished. That would give a clearer picture of how negative or positive performance stimuli persist after aural information has ceased. It could also begin to suggest how recall is affected by recency in an evaluatory context.

No matter what direction future investigations take, it is clear that the examination of how people perceive music in time is a vital, important, and salient area for research. Much of music appreciation is simple discrimination, and the value of “knowing what you like” cannot be understated. At the same time, it is clear that we don’t always know

what we like, or that what we like may be subject to drastic alteration by relatively minor events. Only by continuing to scrutinize the relationship of music to time can we hope to shed further light on this conundrum. This study, while not definitive, provides impetus for that sort of examination, and extends the literature in this area. Further research in these topics is encouraged so that the musical of experience of performers, teachers, and listeners can be optimized.

APPENDIX A

IRB APPROVAL & PARTICIPANT CONSENT FORM

Office of the Vice President For Research
Human Subjects Committee
Tallahassee, Florida 32306-2742
(850) 644-8673 Â· FAX (850) 644-4392

APPROVAL MEMORANDUM

Date: 2/26/2009

To: Robert Simpson [res06e@fsu.edu]

Address: 518 Patty Lynn Drive, Tallahassee, FL 32305
Dept.: MUSIC SCHOOL

From: Thomas L. Jacobson, Chair

Re: Use of Human Subjects in Research
The Amenability of Musical Decisions

The application that you submitted to this office in regard to the use of human subjects in the proposal referenced above have been reviewed by the Secretary, the Chair, and two members of the Human Subjects Committee. Your project is determined to be Expedited per 45 CFR Â§ 46.110(7) and has been approved by an expedited review process.

The Human Subjects Committee has not evaluated your proposal for scientific merit, except to weigh the risk to the human participants and the aspects of the proposal related to potential risk and benefit. This approval does not replace any departmental or other approvals, which may be required.

If you submitted a proposed consent form with your application, the approved stamped consent form is attached to this approval notice. Only the stamped version of the consent form may be used in recruiting research subjects.

If the project has not been completed by 2/25/2010 you must request a renewal of approval for continuation of the project. As a courtesy, a renewal notice will be sent to you prior to your expiration date; however, it is your responsibility as the Principal Investigator to timely request renewal of your approval from the Committee.

You are advised that any change in protocol for this project must be reviewed and approved by the Committee prior to implementation of the proposed change in the protocol. A protocol change/amendment form is required to be submitted for approval by the Committee. In addition, federal regulations require that the Principal Investigator

promptly report, in writing any unanticipated problems or adverse events involving risks to research subjects or others.

By copy of this memorandum, the Chair of your department and/or your major professor is reminded that he/she is responsible for being informed concerning research projects involving human subjects in the department, and should review protocols as often as needed to insure that the project is being conducted in compliance with our institution and with DHHS regulations.

This institution has an Assurance on file with the Office for Human Research Protection. The Assurance Number is IRB00000446.

Cc: Clifford Madsen, Advisor [cmadsen@fsu.edu]
HSC No. 2008.1903

Participant Consent Form
The Amenability of Musical Decisions

You are invited to be in a research study of the perception of musical performance. You were selected as a possible participant because you are trained as a musician, and are at least 18 years of age. Please read this form and ask any questions you may have before agreeing to be in the study.

This study is being conducted by R. Eric Simpson, FSU College of Music.

Background Information:

The purpose of this study is: to examine the amenability of musical decisions. This study will require that you listen to several recordings of a musical performance and indicate your preference for the performance recordings.

Procedures:

If you agree to be in this study, you will be asked to do the following things:

1. Listen to an audiotape of musical excerpts for approximately ten minutes.
2. Indicate your preference for the musical excerpts while they occur using an electronic dial.
3. Indicate your level of musical training.
4. Share your thoughts on this study.

Risks and benefits of being in the Study:

There are no foreseeable risks to participating in this study. There are no individual benefits to participating in this study.

Confidentiality:

The records of this study will be kept private and confidential to the extent permitted by law. In any sort of report that may be published, no information will be included that will make it possible to identify a subject. Research records will be stored securely and only the researcher will have access to the records.

Voluntary Nature of the Study:

Participation in this study is voluntary. Your decision whether or not to participate will not affect your current or future relations with the University. If you decide to

FSU Human Subjects Committee Approved on 2/26/2009. Void after 2/25/2010. HSC#: 2008.1903

participate, you are free to not answer any question or withdraw at any time without affecting those relationships. If you withdraw before the experiment is completed, your results will be completely removed from the study.

Contacts and Questions:

The researcher conducting this study is R. Eric Simpson. You may ask any question you have now. If you have a question later, you are encouraged to contact him at 850.644.6042, or at res06e@fsu.edu. Mr. Simpson's major professor is Clifford Madsen, who may be reached at 850.644.3554.

If you have any questions or concerns regarding this study and would like to talk to someone other than the researcher, you are encouraged to contact the FSU IRB at 2010 Levy Street, Research Building B, Suite 276, Tallahassee, FL 32306-2742, or 850-644-8633, or by email at humansubjects@magnet.fsu.edu.

Statement of Consent:

I have read the above information. I have asked questions and have received answers. I consent to participate in the study.

_____ _____
Signature Date

_____ _____
Signature of Investigator Date

FSU Human Subjects Committee Approved on 2/26/2009. Void after 2/25/2010. HSC#: 2008.1903

APPENDIX B
EXPERIMENTER INSTRUCTIONS

EXPERIMENTER INSTRUCTIONS

1. Distribute participant consent forms and demographic forms. Instruct subjects to complete both forms.
2. Turn on computer.
3. Turn on CD player. Load whichever Order CD that you need.
4. On the computer, double click on “CRDI Alias” to open the CRDI software.
5. Go to the Input Menu and click Select. Choose the number of dials that you will actually be using for this session (1-4). Click on OK. The number of dials that you are using will appear on your screen.
6. Go to the Configure menu. Use “Input/Load saved settings” to load the dial calibration.
7. Click on File and Open the Control File for this project.
8. Read the instructions to the subjects.
9. Click on Run and then on Collect Data button. You may also simply click “Collect Data” underneath the Zones area.
10. Click Start simultaneously when you start the sound source. You are now collecting data.

Saving Data after a Run

1. When data collection is finished, the session data window will be open. Click on View SPI. Click on Save in this Format.
2. Give a name for the data file. Click Save and OK.
3. Observe the check mark next to the Save in SPI Format before clicking Close
4. Click Discard Data and Yes.

Collect Participant Consent Forms and Demographic Forms. Repeat for each group of subjects.

APPENDIX C

DEMOGRAPHIC INFORMATION FORM

Simpson Perception Study - Demographic Information

Name:_____

(Your name is used only for identification purposes, and will be removed from the study once the following demographic data is compiled.)

Years of Musical Training:_____

I am a(n): ☐ Undergraduate ☐ Graduate

Year in program/college:_____

My major instrument is: ☐ String ☐ Vocal ☐ Wind/Perc.

My age is:_____

Free Response Section. Please feel free to place any comments/suggestions/questions about this study here. You may complete this section before or after the study._____

_____

_____

_____

_____

APPENDIX D
INSTRUCTIONS TO SUBJECTS

INSTRUCTIONS

Please read and sign the “participant consent form.” Then, please complete the “participant demographic form.” You may write items in the free-response section now, or wait until the conclusion of the study.

In a moment, you will hear five different performances of the same composition. It will be played on solo cello, and will last for 1 minute and 21 seconds.

As you listen to each performance, please turn the dial in front of you to indicate your perception of the QUALITY OF THAT PEFORMANCE.

Before beginning each listening task, make sure your dial is in the neutral position. There will be a short break (14 seconds) between each performance during which music unrelated to the study will play. Use this time to return your dial to the neutral position.

Please make sure your dial is in the neutral position before we begin.

(AFTER STUDY) At this time, you may complete the free response section of the demographic form if you wish to do so, or you may exit.

APPENDIX E
ANSWERS TO FREE RESPONSE SECTION

Omitted subjects did not respond to this section of the demographic form, or responded in an irrelevant fashion.

Subject 3

What is the purpose of this study? Is the recording for an audition, for professional use, to be sold?

Subject 4

Yes, wrong notes are very unpleasant.

Subject 6

Very interesting. I found myself being more critical with each example.

Subject 10

The “quality of performance” is very relative as you may have intended it to be.

Subject 12

Fascinating study!

Subject 16

I was a little unsure about when to turn the dial or if it could be moved more than once.

Subject 17

I have no problem with you leaving my information out if it skews your statistics. It was hard to make an evaluation between the 1st and 2nd examples because I had nothing to compare them to. The latter ones were judged in comparison to the 1st two.

Subject 26

I found myself thinking more positively of the performances that had a strong (& correct) finish, even if there were mistakes at the beginning – and less of the performances had mistakes throughout.

Subject 27

I usually think that conviction makes a performance better than all the right notes . . . but wrong notes really turned my dial to the negative!

Subject 32

The recordings with the most diverse dynamics were the most pleasant. They seemed to be the most musical. Those with mistakes or odd notes were less enjoyable.

Subject 33

It is amazing how you can hear something over and over and have different perceptions each time, even in the parts I felt were in tune.

Subject 34

The contrast in Baroque music to Renaissance was intriguing. The cello reminded me of Casals & the in-between music of the Quest for the Holy Grail.

Subject 39

It might be nice to have the music in front of the participant. The first example had intonation issues, but I was slightly unsure if they were simply notes that the composer (Bach) had intended.

Subject 40

I may have answered beginning ones differently had I known that quality/preference would involve so many wrong notes! I'm generally forgiving of a few wrong notes, but after awhile it gets increasingly annoying . . . possibly why my later responses were so much more negative than my earlier ones.

Subject 41

I began basing my judgment off of technique, tone quality, etc. and tended to forgive "wrong notes" at first. As the study continued and I realized that the skill level of the performer was relatively inconsistent and the issues were note accuracy and interpretation, I became more critical as time elapsed.

Subject 45

Not giving anything to base the first response to, everything else is compared to that one, no matter if the first one (is) better or worse.

Subject 52

Cool study!

Subject 55

That was interesting, I guess. It seemed like the same performance except with progressively more wrong notes. Hope you find out interesting stuff from this study!

Subject 56

It is quite difficult to translate our musical opinion into a positive/negative wheel. There are more “continuums” in my mind judging music.

Subject 57

I felt it was interesting to see how obvious the errors (or dissonances) in the performances were to me. I guess I don’t tend to notice so much when I’m not asked to rate a performance, usually just on my own playing.

Subject 60

Curious as to what the results will be. Was there any reason for the intermission choice?

Subject 69

I feel like it’s the same person purposefully making mistakes for us to pick up on. In general I thought all the performances went well and the only mistakes made weren’t terrible.

Subject 70

It sounds like the recordings went from great to worse. It progressively kept getting worse with the 2nd and 3rd recording.

Subject 73

I found myself somewhat distracted from the music by my surroundings (papers in front of me, people next to me). Maybe lower the lights and eliminate all other distractions.

Subject 82

I wasn't sure how to judge the notes I was sure the cellist played incorrectly on purpose. They were obviously wrong notes, and some were obviously out-of-tune on purpose. I know this piece very well, so it was obvious to me where the errors were. The only thing I knew for sure is that this is the same cellist playing it five times and that he/she was capable of playing it perfectly because of other factors.

Subject 86

I wish the music in between resolved . . . the first two sounded pretty similar so I probably did something-sorry if I'm an outlier, haha.

Subject 87

Everything sounds the same to me in the first two. The 3rd one sounded pretty bad. 4th wasn't as bad, but there were still some mistakes. The 5th one wasn't the worst.

Subject 88

I was reticent to score the 1st play-through either high or low because I didn't know what to expect or how it would compare to the other play-throughs. But that may very well be part of the experiment.

Subject 89

I wonder which of the performance were actually correct, and how the first performance heard affected the rest of the test. I wonder if the 4th performance was played first, could I have felt negatively about the 1st performance?

Subject 90

I thought I heard some sour notes in the first couple of performances but I wasn't sure.

Subject 91

The amount of wrong notes in the different recordings and where they were placed had a surprising effect on how much I liked/didn't like the recording.

Subject 92

Too many times listening to a cello that constantly sounds worse.

Subject 93

Can we move the dial around and change our mind during a performance or is it once we move it, it has to stay there?

Subject 98

The first listening was "practice" with the dial and not consistent with my other interpretations.

Subject 104

Would help to have one "practice round" to understand activity before it counts.

Subject 105

Solo cello. Quality . . . (always more)-Cool study!

Subject 107

After listening to the same recording consecutive times it becomes boring, but when the "out of tune" parts are played, it brings tension to the piece and re-establishes interest.

Subject 109

It was obviously the same recording with slight pitch variations.

Subject 111

Using the same piece of music was effective for having the participants listen to the quality of the music.

Subject 118

I thoroughly enjoyed the musical palate cleanser.

Subject 119

I think this is a great study! Thank you. Nice Renaissance music.

Subject 124

The fanciful feast music was grand.

Subject 127

Almost same interpretation, missed notes was only variance.

REFERENCES

Abeles, H.F. (1973). Development and validation of a clarinet performance adjudication scale. *Journal of Research in Music Education, 21*(3), 246-255.

Alpert, J. (1982). The effect of disc jockey, peer, and music teacher approval of music on music selection and preference. *Journal of Research in Music Education, 30*(3), 173-186.

Attaignant, P. (2006). Danseries a4 parties: Branle gay. [Various performers]. On *Norton Recorded Anthology of Western Music.* [CD]. New York: Naxos.

Bach, J.S. (1983). Suite for Unaccompanied Cello #3, V. Bourrée. [Yo-Yo Ma]. On *Bach: Unaccompanied Cello Suites.* [CD]. New York: CBS Records Masterworks.

Bach, J.S. (1983). Suite for Unaccompanied Cello #5, IV. Sarabande. [Yo-Yo Ma]. On *Bach: Unaccompanied Cello Suites.* [CD]. New York: CBS Records Masterworks.

Bergee. M.J. (1993). A comparison of faculty, peer and self-evaluation of applied brass jury performances. *Journal of Research in Music Education, 41*(1), 19-27.

Bergee, M.J. (1997). Relationships among faculty, peer, and self-evaulation of applied performances. *Journal of Research in Music Education, 45*(4), 601-612.

Bergee, M. J. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 51*, 342-353.

Bergee, M. J., & McWhirter, J. L. (2005). Selected influences on solo and small-ensemble festival ratings: Replication and extension. *Journal of Research in Music Education, 53*, 177-190.

Bergee, M.J., & Platt, M.C. (2003). Influence of selected variables on solo and small-ensemble festival ratings. *Journal of Research in Music Education, 51*(4), 342-353.

Bigand, E. (1997). Perceiving musical stability: The effect of tonal structure, rhythm, and musical expertise. *Journal of Experimental Psychology: Human Perception and Performance, 23*, 808-822.

Borkenau, P., & Liebler, A. (1993). Convergence of stranger ratings of personality and intelligence with self-ratings, partner ratings, and measured intelligence. *Journal of Personality and Social Psychology, 65*(3), 546-553.

Boyle, J. D., Hosterman, G. L., & Ramsey, D. S. (1981). Factors influencing pop music preferences of young people. *Journal of Research in Music Education, 29*(1), 47-55.

Brand, M., & Burnsed, V. (1981). Music abilities and experiences as predictors of error-detection skill. *Journal of Research in Music Education, 29*(2), 91-96.

Brittin, R.V. (1991). The effect of overtly categorizing music on preference for popular music styles. *Journal of Research in Music Education, 39*(2), 143-151.

Brittin, R.V. (1996). Listeners' preference for music of other cultures: Comparing response modes. *Journal of Research in Music Education, 44*(4), 328-240.

Brittin, R.V., & Duke, R.A. (1997). Continuous and summative evaluation of intensity in music: A comparison of two methods for measuring overall effect. *Journal of Research in Music Education, 43*(4)*,* 36-46.

Brittin, R.V., & Sheldon, D. A. (1995). Comparing continuous versus static measurements in music listeners' preferences. *Journal of Research in Music Education, 43*(1), 36-46.

Bullock, W. J. (1973). A review of measures of musico-aesthetic attitude. *Journal of Research in Music Education, 21*(4), 331-344.

Cassidy, J. W., & Sims, W. L. (1996). Effects of special education labels on peers' and adults' evaluation of a handicapped youth choir. *Journal of Research in Music Education, 39*, 23-34.

Colwell, C.M. (1995). Effect of teaching setting and self-evaluation on teacher intensity behaviors. *Journal of Research in Music Education, 43*(1), 6-21.

Cooksey, J.M. (1977). A facet-factorial approach to rating high-school choral performance. *Journal of Research in Music Education, 25*(2), 100-114.

Crafts, L.W. (1932). Primacy and recency in the learning of visual diagrams. *The American Journal of Psychology. 44*(4), 763-767.

Crowder, R.G. (1976). *Principles of learning and memory.* New York: Halsted Press.

Deal, J.J. (1985). Computer assisted-instruction in pitch and rhythm error detection. *Journal of Research in Music Education, 33*(3), 159-166.

Duerksen, G. L. (1972). Some effects of expectation on evaluation of musical performance. *Journal of Research in Music Education, 20*, 268-272.

Duke, R.A., & Colprit, E.J. (2001). Summarizing listener perceptions over time. *Journal of Research in Music Education, 49*(4), 330-342.

Ebbinnghaus, H. (1964). *Memory: A contribution to experimental psychology* (H.A. Ruger & C.E. Bussenius, Trans.) New York: Dover. (Original work published 1885).

Fiske, H. E. J. (1975). Judge-group differences in the rating of secondary-school trumpet performances. *Journal of Research in Music Education, 23*, 186-196.

Fiske, H. E. J. (1977). Relationship of selected factors in trumpet performance adjudcation reliability. *Journal of Research in Music Education, 25*, 256-263.

Fredrickson, B.L., & Kahneman, D. (1993). Duration neglect in retrospective evaluations of affective episodes. *Journal of Personality and Social Psychology, 65*(1), 45-55.

Fredrickson, W. E., Johnson, C. M., & Robinson, C. R. (1998). The effet of pre-conducting and conducting behaviors on the evaulation of conductor competence. *Journal of Band Research, 33*(2), 1-13.

Fung, V. C. (1996). Musicians' and nonmusicians' preferences for world musics: Relation to musical characteristics and familiarity. *Journal of Research in Music Education, 44*(1), 60-83.

Furman, C. A., & Duke, R. A. (1988). Effect of majority consensus on preferences for recorded orchestral and popular music. *Journal of Research in Music Education, 36*(4), 220-231.

Galdi, S., Arcuri, L., & Gawronski, B. (2008). Automatic mental associations predict future choices of undecided decision-makers. *Science, 321*(5892), 1100-1102.

Geringer, J.M. (1995). Continuous loudness judgments of dynamics in recorded music excerpts. *Journal of Research in Music Education, 43*(1), 22-35.

Geringer, J. M., & Dunnigan, P. (2000). Listener preferences and perception of digital versus analog live concert band recordings. *Bulletin of the Council for Research in Music Education, 145*, 1-13.

Geringer, J. M., & Madsen, C. K. (1981). Verbal and operant discrimination-preference of tone quality and intonation. *Psychology of Music, 9*, 26-30.

Geringer, J.M., & Madsen, C.K. (1984). Pitch and tempo discrimination in recorded orchestral music among musicians and non-musicians. *Journal of Research in Music Education, 32*(3), 195-204.

Geringer, J. M., & Madsen, C. K. (1987). Pitch and tempo preferences in recorded popular music. In C. K. Madsen, & C. A. Prickett (Eds.), *Applications of research in music behavior* (pp. 204). Tuscaloosa, AL: University of Alabama Press.

Geringer, J. M., & Madsen, C. K. (1989). Pitch and tone quality discrimination and preference: Evidence for a hierarchical model of music elements. *Canadian Music Educator, 30*(2), 29-38.

Geringer, J. M., & Madsen, C. K. (1998). Musicians' ratings of good versus bad vocal and string performances. *Journal of Research in Music Education, 46*(4), 522-534.

Goldin, C., & Rouse, C. (2000). Orchestrating impartiality: The impact of "blind" auditions on female musicians. *The American Economic Review, 90*(4), 715-741.

Gotlieb, H., & Konecni, V.J. (1985). The effects of instrumentation, playing style, and structure in the *Goldberg Variations* by Johann Sebastian Bach. *Music Perception, 3,* 87-102.

Gregory, D. (1994). Analysis of listening preferences of high-school and college musicians. *Journal of Research in Music Education, 42*(4), 331-342.

Gregory, D. (1996). Reliability of the Continuous Response Digital Interface. *Proceedings of the Third International Technological Directions in Music Education Conference.* San Antonio, TX: Institute for Music Research.

Halper, A.R., & Bower, G.H. (1982). Musical expertise and melodic structure in memory for musical notation. *The American Journal of Psychology, 95*(1), 31-50.

Hargreaves, D., J. (1984). The effects of repetition on liking for music. *Journal of Research in Music Education, 32*(1), 35-47.

Hewitt, M. P., & Smith, B. P. (2004). The influence of teaching-career level and primary performance instrument on the assessment of music performance. *Journal of Research in Music Education, 52*(4), 314-327.

Jenkins, J.G. (1927). The effect of serial position on recall. *The American Journal of Psychology, 38*(2), 285-291.

Johnson, C. M., & Stewart, E. E. (2005). Effect of sex and race identification on instrument assignment by music educators. *Journal of Research in Music Education, 53*, 248-257.

Juchniewicz, J. (in press). Listener expectation on the evaulation of musical performance. *Southeastern Journal of Music Education,*

Juchniewicz, J. (2005). *The influence of physical movement on the perception of musical performance.* Unpublished Master's Thesis, Florida State University, Tallahassee, FL.

Kahneman, D. & Snell, J. (1992). Predicting a changing taste: Do people know what they will like? *Journal of Behavioral Decision Making, 5*(3), 187-200.

Kahneman, D., Fredrickson, B., Schreiber, C., & Redelmeier, D.A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 400-405.

Killian, J.N. (1987). The effect of choral compositional style on operant balance preference. In C. K. Madsen, & C. A. Prickett (Eds.), *Applications of research in music behavior* (pp. 213). Tuscaloosa, AL: University of Alabama Press.

Killian, J.N. (1990). Effects of model characteristics on musical preference of junior high students. *Journal of Research in Music Education, 38*, 115-123.

Killian, J.N. (1991). The relationship between sightsinging accuracy and error-detection in junior-high singers. *Journal of Research in Music Education, 39*(3), 316-324.

Konecni, V.J. (1984). Elusive effects of artists' "messages." In W.R. Crozier, & A.J. Chapman (Eds.), *Cognitive Processes in the Perception of Art.* Amsterdam: Elsevier Science Publishers, B.V.

Kuhn, T. L. (1987). The effect of tempo, meter, and melodic complexity on the perception of tempo. In C. K. Madsen, & C. A. Prickett (Eds.), *Applications of research in music behavior* (pp. 165). Tuscaloosa, AL: University of Alabama Press.

Langer, S.K. (1957). *Philosophy in a new key: A study in the symbolism of reason, rite, and art.* Cambridge, MA: Harvard University Press.

Larsen, R.C. (1977). Relationships between melodic error-detection, melodic dictation, and melodic sightsinging. *Journal of Research in Music Education, 25*(4), 264-271.

Leblanc, A. (1980). Outline of a proposed model of sources of variation in musical taste. *Bulletin of the Council for Research in Music Education, 61*, 29-34.

Leblanc, A., & Cote, R. (1983). Effects of tempo and performing medium on children's music preference. *Journal of Research in Music Education, 31*(4), 283-294.

Leblanc, A., Sims, W. L., Siivola, C., & Obert, M. (1996). Music style preferences of different age listeners. *Journal of Research in Music Education, 44*(1), 49-59.

Levinson, J.L. (1997). *Music in the moment.* Ithaca, NY: Cornell University Press.

Madsen, C.K. (1997). Focus of attention and aesthetic response. *Journal of Research in Music Education, 45*(1), 80-89.

Madsen, C.K., Brittin, R.V., & Capperella-Sheldon, D.A. (1993). An empirical method for measuring the aesthetic response to music. *Journal of Research in Music Education, 41*(1), 57-69.

Madsen, C.K., Duke, R.A., & Geringer, J.M. (1986). The effect of speed alterations on tempo note selection. *Journal of Research in Music Education, 34*(2), 101-110.

Madsen, C.K., & Fredrickson, W.E. (1993). The experience of musical tension: A replication of Neilsen's research using the continuous response digital interface. *Journal of Music Therapy, 30*(1), 46-63.

Madsen, C. K., & Geringer, J.M. (1976). Preferences for trumpet tone quality versus intonation. *Bulletin of the Council for Research in Music Education, 46*, 13-22.

Madsen, C. K., Geringer, J.M., & Wagner, M. J. (2007). Context specificity in music perception of musicians. *Psychology of Music, 35*, 441-451.

Madsen, C. K., & Moore, R. S. (1978). *Experimental research in music: Workbook in design and statistical tests* (2nd ed.). Raleigh, NC: Contemporary Publishing Company of Raleigh, Inc.

Madsen, C. K., & Prickett, C. A. (Eds.). (1987). *Applications of research in music behavior*. Tuscaloosa, Alabama: University of Alabama Press.

Maess, B., Koelsch, S., Gunter, T.C., & Friederici, A.D. (2001). Musical syntax is processed in Broca's area: An MEG study. *Nature Neuroscience, 4*, 540-545.

McCrary, J. (1993). Effects of listeners' and performers' race on music preferences. *Journal of Research in Music Education, 41*(3), 200-211.

McRary, J.W., & Hunter, W.S. (1953). Serial position curves in verbal learning. *Science, 117*(3032), 131-134.

Meyer, L.B. (1956). *Emotion and meaning in music.* Chicago: University of Chicago Press.

North, A. C., & Hargreaves, D. J. (1997). Liking, arousal potential, and the emotions expressed by music. *Scandanavian Journal of Psychology, 38*(1), 45-53.

Nosek, B.A., Banaji, M.R., & Greenwald, A.G. (n.d.) *Project Implicit.* Retrieved August 1, 2008 from https://implicit.harvard.edu/implicit/.

Pembrook, R.G. (1987). The effect of vocalization on melodic memory conservation. *Journal of Research in Music Education, 35*(3), 155-169.

Plack, D.S. (2006). The effect of performance medium on the emotional response of the listener as measured by the Continuous Response Digital Interface. Unpublished Doctoral Dissertation, Florida State University, Tallahassee, FL.

Pro Tools 8 HD (2008). [Computer software]. Menlo Park, CA: Digidesign.

Radocy, R. E. (1976). Effects of authority figure biases on changing judgments of musical events. *Journal of Research in Music Education, 24*(3), 119-128.

Radocy, R. E., & Boyle, J. D. (2003). *Psychological foundations of music behavior* (2nd ed.). Springfield, Illinois: Charles C Thomas.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior, 6*, 855-863.

Rentfrow, P. J., & Gosling, S. D. (2003). The do re mi's of everyday life: The structure and personality correlates of music preferences. *Journal of Personality and Social Psychology, 84*, 1236-1256.

Rentz, E. (1992). Musicians' and non-musicians' aural perception of orchestral instrument families. *Journal of Research in Music Education, 40*(3), 185-192.

Robinson, C.R. (1988). Differentiated modes of choral performance evaluation using traditional procedures and a Continuous Response Digital Interface. Unpublished doctoral dissertation, Florida State University, Tallahassee, FL.

Rohrer, T. P. (2002). The debate on competition in music in the twentieth century. *Update: The Applications of Research in Music Education, 21*, 17-25.

Rozin, A., Rozin, P., & Goldberg, E. (2004). The feeling of music past: How listeners remember musical affect. *Music Perception, 22*(1), 15-39.

Ryan, C., & Costa-Giomi, E. (2004). Attractiveness bias in the evaluation of young pianists' performances. *Journal of Research in Music Education, 52*(2), 141-154.

Ryan, C., Wapnick, J., Lacaille, N., & Darrow, A. A. (2006). The effects of various physical characteristics of high-level performers on adjudicators' performance ratings. *Psychology of Music, 34*(4), 559-572.

Seashore, C.E. (1908). Elementary experiments in psychology. New York: Holt.

Shehan, P. K. (1985). Transfer of preference from taught to untaught pieces of non-western music genres. *Journal of Research in Music Education, 33*(3), 149-158.

Sheldon, D.A. (1994). Effect of tempo, musical experience, and listening modes on tempo modulation perception. *Journal of Research in Music Education, 42*(3), 190-202.

Sheldon, D.A. (2004). Effects of multiple listening on error-detection acuity in multivoice, multitimbral musical examples. *Journal of Research in Music Education, 52*(2), 102-115.

Sloboda, J.A. (1976). Visual perception of musical notation: Registering pitch symbols in memory. *The Quarterly Journal of Experimental Psychology, 28*(1), 1-16.

Southall, J.K. (2003). The effect of purposeful distractors placed in an excerpt of Puccini's *La Bohème* to ascertain their influence on the listening experience. Unpublished Doctoral Dissertation, Florida State University, Tallahassee, FL.

Tillman, B., Bharucha, J.J., & Bigand, E. (2000). Implicit learning of tonality: A self-organizing approach. *Psychological Review, 107*, 885-913.

Tillman, B., & Bigand, E. (2004). The relative importance of local and global structures in music perception. *The Journal of Aesthetics and Art Criticism, 62*(2), 211-222.

Thompson, S., & Williamon, A. (2003). Evaluating evaluation: Musical performance assessment as a research tool. *Music Perception, 21*, 21-41.

Thorndike, E.L. (1927). The influence of primacy. *Journal of Experimental Psychology, 10,* 18-29.

VanWeelden, K. (2002). Relationships between perceptions of conductor effectiveness and ensemble performance. *Journal of Research in Music Education, 50*(2), 165-176.

VanWeelden, K., & McGee, I. R. (2007). The influence of music style and conductor race on perceptions of ensemble and conductor performance. *International Journal of Music Education, 25*, 7-17.

Wapnick, J. (1976). A review of research on attitude and preference. *Bulletin of the Council for Research in Music Education, 48*, 1-20.

Wapnick, J., & Rosenquist, M.J. (1987). Preferences of undergraduate music majors for sequenced versus performed piano music. *Journal of Research in Music Education, 39*(2), 152-160.

Wapnick, J., Darrow, A. A., Kovacs, J., & Dalrymple, L. (1997). Effects of physical attractiveness on evaluation of vocal performances. *Journal of Research in Music Education, 45*, 470-479.

Wapnick, J., Flowers, P., Alegant, M., & Jasinkas, L. (1993). Consistency in piano performance evaluation. *Journal of Research in Music Education, 41*, 282-292.

Wapnick, J., Kovacs, J., & Darrow, A. A. (1998). Effects of performer attractiveness, stage behavior, and dress on violin performance. *Journal of Research in Music Education, 46*(4), 510-521.

Wapnick, J., Mazza, J. K., & Darrow, A. A. (2000). Effect of performer attractiveness, stage behavior, and dress on evaluation of children's piano performances. *Journal of Research in Music Education, 48*(4), 323-335.

Wapnick, J., Ryan, C., Campbell, L., Deek, P., Lemire, R., & Darrow, A. A. (2005). Effect of excerpt tempo and duration on musicians' ratings of high-level piano performances. *Journal of Research in Music Education, 53*(2), 162-176.

Williams, D. A. (1996). Competition and music - who are the winners? *Update: The Applications of Research in Music Education, 15*, 16-21.

Williams, D.B. (1975). Short-term retention of pitch sequence. *Journal of Research in Music Education, 23*(1), 53-66.

Welch, G.B., & Burnett, C.T. (1924). Is primacy a factor in association-formation? *The American Journal of Psychology, 35*(3), 396-401.

Yarbrough, C. (1987). The effect of musical excerpts on tempo discriminations and preferences on musicians and non-musicians. In C. K. Madsen, & C. A. Prickett (Eds.), *Applications of research in music behavior* (pp. 175). Tuscaloosa, AL: University of Alabama Press.

Yarbrough, C., Wapnick, J.L., & Kelly, R. (1976). Effect of videotape feedback on performance, verbalization and attitude of beginning conductors. *Journal of Research in Music Education, 27*(2), 103-112.

BIOGRAPHICAL SKETCH

Name:	Robert Eric Simpson
Birthplace:	Ocala, Florida
Date of Birth:	September 12, 1975
Higher Education:	Stetson University DeLand, FL Major: Music Education Degree: B.M.E. (1997)
	University of Cincinnati College-Conservatory of Music Cincinnati, OH Major: Music Education Degree: M.M. (2006)
	Florida State University Tallahassee, FL Major: Music Education Degree: Ph.D. (2009)
Professional Experience:	Fort King Middle School Ocala, FL 1997-1998 Director of Bands
	Dr. Phillips High School Orlando, FL 1998-2000 Associate Director of Bands

Professional Experience (Cont.):	William R. Boone High School Orlando, FL 2000-2004 Director of Bands; Chair, Visual and Performing Arts
	Texas Christian University Fort Worth, TX Fall 2009 Assistant Professor, Music Education