CUE WEIGHTING AND TASK PROCESSING IN MELODIC CLASSIFICATION

Duncan Snape
Master's Thesis
Music, Mind, and Technology
Department of Music
5th May 2015
University of Jyväskylä

JYVÄSKYLÄN YLIOPISTO

Tiedekunta – Faculty Humanities	Laitos – Department Music Department
Tekijä – Author Duncan Snape	
Työn nimi – Title Cue Weighting and Task processing in Melodic Similarity	
Oppiaine – Subject Music, Mind & Technology	Työn laji – Level Master’s Thesis
Aika – Month and year May 2015	Sivumäärä – Number of pages 88

Tiivistelmä – Abstract

Many studies have considered the role of diverse musical parameters in the perception and cognition of specific musical contexts, yet few attempts have been made to integrate these findings into a general cognitive overview of everyday music listening. This study shifted the focus of research, by considering the importance, the relative weight, of particular parameters of music to individual music listeners, and groups composed of those listeners who had differing levels of formal musical experience. Manipulations in various musical and auditory parameters – melodic contour, metrical accent, instrument timbre, distortion and loudness – in two short melodic phrases, were presented to two groups of participants, the formal and informal group, for melodic similarity classification in a factorial design. Individual participants employed a wide variety of weighting schemas, both in parameter hierarchy and distribution, far more than could be satisfactorily explained by the idea of parametric salience alone. Participant groups employed collective schemas that differed significantly from each other and their perceptual error rates were different both between and within-group for parts of the task where different parameters were available. The formal group exhibited an overall task processing advantage, needing less time to complete each matrix than the informal group. The greatest difference in processing between groups was in the mean excerpt listening time. Informal group participants listened for longer, particularly during secondary groupings.

Asiasanat – Keywords
music cognition, cue abstraction, cue weighting, salience, melodic similarity

Säilytyspaikka – Depository	
Muita tietoja – Additional information	

Acknowledgements

My thanks to those who have gone beyond my expectations during my studies in Jyväskylä to provide advice, feedback and support – Rafael Ferrer, Anemone Van Zijl, Birgitta Burger and of course Markku Pöyhönen.

Special thanks are due to my mentor in the early stages of this study, Professor Justin London, without whose input things might have gone very differently. Also to Professor Stephen Croucher for statistics support, to Geoff Luck for feedback and advice, and to my supervisor Marc Thompson for patience, understanding (third time lucky!), and ice cream with our children on sunny days.

Finally, to my beautiful and talented partner Nerdinga Letulė, my strongest supporter and harshest critic, for help, wisdom, experience and snuggling. And cake.

Table of Contents

1 INTRODUCTION

Listening puts me in the world. Listening gives me a sense of emotion, a sense of movement, and a sense of being there that is missing when I am looking – Stephen Handel

More than any other human activity, music remains one of the most diffuse and indefinable. The perception and cognition of music have occupied scientific inquiry throughout human history, and prominent thinkers have expounded on these topics since the times of the Ancient Greeks. Yet whilst large bodies of knowledge have accrued in more recent times about the physical and neural mechanisms underlying auditory perception, general cognitive processes and the specific processing and understanding of isolated attributes of music, relatively few attempts have been made to piece these strands together into a functional model of the general cognitive strategies employed in everyday music listening. The question pertinent to such an aim is 'how do we understand music while we are listening to it?', and this should, ideally, be answered before we can successfully progress to the next logical step - '…and what does the way we listen reveal to us about the music?' The latter often seems to be the starting point in music cognition research, whilst the former has yet to be fully addressed.

Through influential research from authors such as Lerdahl and Jackendoff (1985), Bregman (1994) and Deliege (1987, 1989), we have plausible explanations for some aspects of how music is understood during listening. We know that the perception of complex auditory stimuli seems to be based on the grouping principles proposed by Gestalt psychology (Bregman, 1994), and that those same grouping principles can be applied to the process of musical sense-making (Lerdahl & Jackendoff, 1985). We also know that the basis for these grouping decisions is a set of abstracted salient features, or cues, which facilitate the construction of a mental schema of the piece of music being listened to (Deliege, 1987; 1989). But the roles of these individual mechanisms in music cognition are still not completely understood nor researched.

Many studies have highlighted the roles of different musical parameters as sources of cues in the perception of diverse musical phenomena such as structural detail (e.g. Bruderer, McKinney & Kohlrausch, 2010), the building and release of tension (e.g. Granot & Eitan, 2011) and the communication of emotion (e.g. Juslin, 2000), each coming to its own conclusions about the relative function and importance of the parameters highlighted. Yet, if we consider a wider principle supposed to underlie much of the processing of sensory input during our daily lives, that of cognitive economy, in relation to music listening (see Reybrouck (2010) for a discussion of cognitive economy in music listening), it seems unlikely, intuitively, that all the proposed perceptual models for the identification of different dimensions of music can function simultaneously and effectively. Indeed Deliege's (2001) summary of the cue abstraction process indicates that it is already one of economy, in which a cue acts as a marker in the musical stream, facilitating recognition without *re-cognition*. It seems, then, that greater attempts need to be made to place these context-specific findings in an overall framework which reflects the wider experience of everyday music listening, and that this may require a change in general approach.

For *our* cue as music researchers on how to proceed in quantifying the real-time, naturalistic perception of music, we submit that we should look both to general psychological research on perception and cognition, and to other strands of research dealing with auditory domains.

Links between the phenomena of music and language are well established in the literature. Recent research and review has highlighted commonalities in perception (e.g. Besson, Chobert & Marie, 2011), syntax (e.g. Patel, 2007) and neural processing (e.g. Levitin & Menon, 2003). Patel (2007), in particular, makes much of the use of hierarchical structures in the syntactic organisation of both language and music, stating that: "These similarities are interesting because they suggest basic principles of syntactic organization employed by the human mind." (p.267).

Narrowing the focus of our inquiry to purely linguistic research, we encounter evidence of another potentially interesting hierarchical perceptual structure in the work of Holt and Lotto. In their 2006 study, the authors both summarize, and contribute to, a body of research concerned with the phenomenon of cue weighting in speech perception. Cue weighting refers to the observation that language listeners show preferential bias towards certain acoustic cue sources in categorical perception, the process of identifying a sound as belonging to a family of other sounds. In the aforementioned study, participants were trained to categorize simple sine waves based on manipulations along two dimensions, center frequency and modulation frequency. Participants showed a bias for center frequency as the basis for discrimination, even when stimuli were manipulated to make the preferred parameter less informative as a cue source. The results of some of the context-specific musical studies already highlighted suggest that similar processes might be at work in music-listening, but no studies are apparent in the music literature that have explicitly explored such a possibility.

Another potentially interesting dimension to the cue weighting phenomenon, as presented in Lotto and Holt (2006), is that of its underlying causes. The authors suggest a strong link between cue weighting and experiential bias, that is, the perception of a cue source as being more informative as a result of experiential confirmation. This stands in marked contrast to the proposed mechanism, both implicit and explicit in music research, at the root of cue abstraction – that it occurs purely as a function of the salience of a given musical parameter in context (Deliege, 1996). Individual experiential bias might provide a better explanation for the variation in individual responses frequently observed in music perception research (see, for example, Juslin, 2000; Eerola, Järvinen, Louhivuori & Toiviainen, 2001; McAdams, Vieillard, Houix & Reynolds, 2004; Bruderer, McKinney & Kohlrausch, 2010) than parameter salience alone, and would also suggest a greater role of subjectivity in music perception/cognition than previously considered in the literature.

The scope of this thesis is to attempt a first step towards a more general answer to the question of how we understand music while listening to it. It will do this by investigating the concept that a weighting of acoustic cues occurs in the cognitive processing of musical materials. Specifically, this investigation will be applied to a wider listening context analogous to a typical Western listener's daily interaction with music. It will shift the emphasis of experimentation away from trying to determine which musical parameters contribute the most to the perception of specific musical devices, to focus on determining the relative weight of these parameters to individual music listeners. It will also explore what differences are manifest in the weighting of parameters and processing of materials between groups comprised of individuals with differing levels of formal musical experience.

The following general research questions will be addressed:

- Do participants show evidence of individual cue weighting in the use of parameters available for musical similarity judgements?
- If cue weighting occurs does it relate more closely to salience or bias?
- What differences in parameter selection, task processing and performance are exhibited by participants with differing levels of formal musical training?

The subsequent work is structured as follows:
In the Theoretical Background section, contributions from the literature on similarity, cues, cue abstraction, cue weighting and listening are discussed in more detail, supported and exemplified by the findings of relevant research. In the Method section, the experimental materials and procedures used in this study are thoroughly defined and documented. In the Results section, the data gathered during the experimental procedure is reported and statistical analysis is presented. In the Discussion section, the implications of the results are discussed and considered in the context of current knowledge. Finally, in the Conclusion, the key findings are summarised, the limitations of the study are considered, and recommendations for future research are proposed.

2 THEORETICAL BACKGROUND

2.1 Similarity, Classification, Cue Abstraction, Cue Weighting

Fundamental to our understanding of the world around us is the ability to break down sensory input into smaller and more easily-digestible pieces of information, in order to compare these pieces to examples stored in long and short-term memory. Reybrouck (2010) describes this procedure, in real-time music listening, as “a dynamic tension between ‘experience’ and ‘recognition’ with the former relying on a moment-to-moment scanning of sensory particulars, and the latter relying on processes of abstraction and generalisation.” (p.188). One of the most ubiquitous of those processes of generalisation is the similarity/dissimilarity judgement.

Similarity judgements facilitate listener recognition of an auditory event and its comparison to previous experience (Reybrouck, 2009, p.112). Similarity is considered the foundation of widely employed structural forms in music – thematic and motivic identity, repetition and variation (Deliege, 2001, p233). Similarity judgements also allow listeners to group events such as discrete pitches and timbres together into classes, both sequentially and simultaneously (Bregman, 1994). It should be noted that whilst similarity is the basis of one of the grouping principles of Gestalt psychology, it is also, simultaneously, the basis of all grouping principles, since all acts of classification and/or categorization require a decision as to whether objects or events belong together (whether they are similar) or not (whether they are dissimilar), an idea discussed by both Deliege (2001, p.235) and Handel (1993, p187-189), and utilised in the development of models of real-time music listening (Deliege, 1989).

Several recent studies have used listeners’ subjective similarity judgements to explore the perception of thematic and motivic materials. Eerola et al. (2001) extracted statistical information about the frequency, metrical and durational properties of 15 folk melodies and used multiple regression to determine how much these properties had contributed to listener’s similarity ratings of the materials. Lamont and Dibben (2001) used similarity ratings of pairs of extracts of piano pieces from Beethoven and Schoenberg, coupled with adjective ratings of the paired stimuli, to explore which musical parameters listeners had used as the basis of similarity judgement. McAdams et al. (2004) asked participants to judge the similarity of 34 excerpts from a piece of contemporary music, which were accompanied by free response descriptions of the reasons for decisions. Ziv and Eitan (2007) used the same stimuli as Lamont and Dibben (2001) and compared similarity ratings collected in the former study to categorizations of belonging made by participants in their experiment. Eitan and Granot (2009) used controlled artificial stimuli, which participants grouped according to what seemed to belong together, in order to determine whether listeners used the parameters suggested by music theory or by more general auditory perception in grouping decisions.

The assumption underlying all these studies is that musical parameters present cues, denoting the onset and cessation of auditory events, to listeners. Cues are salient components, prominent at the musical surface (Deliege, 2001), that function as reference points for comparative strategies of musical sense-making (Deliege, 1989). Cues thus form the basis of a mental schema of a musical work that is stored in the memory and elaborated upon with subsequent listens (Deliege, 1989). The process of recognising and cataloguing cues has been formalized as cue abstraction (Deliege, 1996).

Some authors have set out to explore the relative contribution of diverse musical parameters, as cue sources, to the perception of distinct musical mechanisms. Participants in Bruderer, McKinney and Kohlrausch (2010) were asked to segment both polyphonic midi and polyphonic audio versions of popular songs at perceived structural boundaries during repeated real-time listens. The salience of the boundary in each segmentation was rated and described with free responses, which were grouped into cue classes. The most frequently mentioned cues denoting perceived structural boundaries were harmonic progression, change in rhythm, change in timbre and change in tempo. Granot and Eitan (2011) investigated the interaction of musical parameters in the perception of musical tension. Groups of participants rated a series of short melodic phrases, in which four parameters – pitch contour, tempo, dynamics and pitch register were systematically manipulated, for overall perceived tension and tension change. Dynamics and pitch register emerged as the strongest factors in determining musical tension, but their results showed that parameters interacted in interesting and sometimes unexpected ways. For example, the effects of pitch contour were modulated by pitch register – rising pitch was considered more tense than falling pitch in higher registers, but the same or less tense in lower registers. Importantly, it was found that listeners employed different parameters in rating overall tension than they did in rating tension change, and that parameter choices also differed between musician and non-musician participants.

Juslin (2000) asked participants to rate the emotion of excerpts taken from electric guitar performances of well-known melodies on four adjective scales: happy, sad, angry and fearful. These ratings were compared to the variation in the performances in five defined cue sources according to the performers' expressive intentions. The five cue sources were mean tempo, mean sound level, frequency spectrum, mean articulation, and articulation variability. A number of findings indicated that i) performers were generally successful in conveying intended emotions, ii) different emotions were associated with different combinations of cues, iii) different performers varied in their ability to convey intended emotions to listeners because iv) there were systematic differences between the cue utilization of performers and listeners, but v) two performers who utilized different communicative cues conveyed emotions equally well to listeners. It was further acknowledged that the interaction of melodic structure in the excerpts had had some effect on the ability of performers to convey the required emotions, a factor that had not been controlled for.

Reviewing the identified studies on similarity and cue use it is apparent that the concept that participants ascribe different weights to different sources of perceptual cues is not a novel one – it is clearly, if not explicitly, implied in both the design and results of all but one (Ziv & Eitan, 2007). However, because this concept has yet to be explicitly discussed in relation to music there is no definition nor understanding of what this might imply - to quote Juslin (2000), in relation to the observed differences between his performers and listeners:

> The reasons for these differences are not known. However, it is possible that some of the differences reflect differences between performers and listeners with respect to their expertise. For example, more expertise is probably needed to appreciate differences in articulation than differences in tempo. (p.1809)

In all the music-related literature discussed so far, both theoretical and experimental, a common underlying assumption seems to be that the reason for the unequal weighting of cues is the salience of the source, i.e. whichever parametric component emerges as the most salient - the most noticeable or prominent – it will be this one that is used as the basis for cues. Deliege herself states that cues can be "thematic or rhythmic motifs, intervals, physico-acoustic [*sic*] characteristics (layout of textural densities, timbres, registers, dynamics) etc.; only their specific pertinence in the course of listening will decide" (1996, p.134). Indeed in Eerola at al. (2001) this assumption was explored to such an extent that listener perception of the cues sources leading to their similarity judgements was not collected. In research that has formalized the phenomenon of cue weighting, however, the pre-eminence of salience as the basis of groupings is not an automatic assumption.

In the visual domain, the consensus is that the weight attributed to a cue is in proportion to its reliability as a source of information (e.g. Jacobs, 1999; Landy & Kojima, 2001; Ernst & Banks, 2002). However, as Toscano and McMurray (2010) point out, weighting by reliability, which works well in the visual domain due to the relatively linear relationships between cue sources and their variance, cannot be applied as simply to other perceptual situations where the same circumstances do not exist. In speech perception, for instance, while acoustic cues can be considered continuous: “their statistical distributions are shaped into clusters of cue values by the phonological categories of the language. The listener’s goal is to determine the underlying phonological category from these cues, not necessarily a continuous estimate” (p.435). Music and language utilise those same continuous acoustic cues, and, we submit, the ‘phonology’ of music is also one of category, in the way it treats everything from pitch class, to structure to timbre, presenting the same kind of perceptual problems for music as speech.

According to Holt and Lotto (2006) it is usually necessary, in determining the categorical membership of a sound, to integrate multiple acoustic dimensions. Research has shown, however, that not all dimensions are perceptually equivalent. They cite studies by Hillenbrand, Clark and Houde (2000) and Francis, Baldwin and Nusbaum (2000) that demonstrated perceptual bias for one parameter over another in experimental settings, as well as presenting their own experimental findings where this bias was highlighted again. The authors also submit a set of arguments, based on observations from linguistic research, as to why the weighting of cues occurs. These are, in summary; “as a result of experience with regularities in the input, the robustness or variability of the perceptual coding of an acoustic dimension, and its informativeness to category identity as a function of task” (p.3061). This argumentation is further qualified by the following:

> An appreciation of perceptual cue weighting leads to the perspective that speech categorization is not just a matter of detecting available auditory cues along various acoustic dimensions, but also applying some weighting function that is, at least in part, dependent on experience with phonetic distributions. (p.3061)

The proposal, then, is that listener experience of the sonic environment can lead to listener bias in the selection of an auditory parameter as a source of cues.

Evidence of the process of listener experience shaping perceptual bias is offered by Kuhl (2004), who presents a review of research pertaining to language acquisition in infants. In it the concept of neural commitment emerges. In essence, research has shown that infants are born with the capacity to discriminate between phonetic contrasts in all languages, but within the first 12 months of their lives this capacity decreases in favour of the specific phonology of the language that they most commonly experience. This in turn leads to a greater capacity to learn within the specific and related phonologies, the trade-off being a proportionate decrease in the capacity to learn other phonologies.

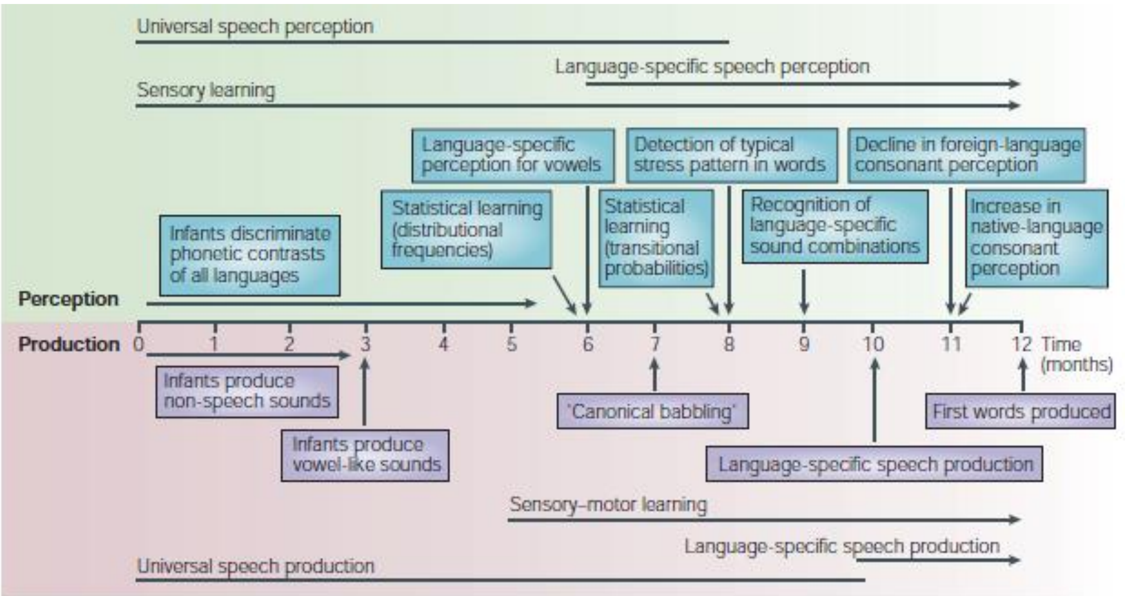


Figure 1. The universal language timeline of speech-perception and speech-production development. (Reproduced from Kuhl, 2004)

Neural commitment thus exemplifies a kind of perceptual specialization that better equips the brain to deal with the stimuli which it commonly experiences, and can offer a potential explanation for the idea of a perceptual bias. Similar specialization processes have been considered in music research. Balkwill and Thompson (1999) proposed a model of emotion perception in music featuring both cultural and psychophysical cues – features (tempo, melodic complexity, rhythmic complexity, pitch range and timbre) common to all music regardless of tonal system. Their results indicated that Western listeners were able to successfully identify the emotions of joy, anger and sadness in Indian classical music from psychophysical cue sources, but were not able to identify peacefulness, the cues for which were considered more culture specific.

The arguments presented by Holt and Lotto (2006) with regard to experiential bias and neural encoding, modulated in some measure by task-specific plasticity, would seem to offer a more robust explanation of the between-subject variability observed in much of the music literature discussed here. Both the assessment of cues in similarity judgements and those used in the perception of common musical devices might conceivably be affected more by experience than relative cue salience. Such a proposal would also lead to questioning the validity of a purely statistical approach to cue weighting, based on the concept of salience, such as that investigated in Eerola et al. (2001). As the authors themselves discussed (p.286), it simply may not account particularly well for perceptual responses.

It seems, then, that cue weighting may be affected at an individual level by experience, making the choice of a source of cues in a less task-specific context more subjective than has previously been considered in music research. Batt-Rawden and De Nora (2005) and Folkestad (2006) have both argued that daily exposure to music constitutes an informal learning environment. As we experience and learn language, behaviours and social convention from those around us, so we also experience and learn music. We learn how it is formed and therefore what we expect it to sound like. We also learn what functions it fulfils and, importantly, whether or not we like it. Since music does not exhibit the same fixed set of lexical units of meaning as language, it follows that without the moderating influence of formal training an individual's understanding of music may emerge in relatively idiosyncratic ways, which depend to a great extent on the volume and nature of their social and contextual exposure to it. This in turn may lead to interpretations of what constitute the important parameters of music that differ quite considerably from individual to individual. It seems counterproductive, therefore, to start from a position of labelling some listeners as inexperienced or naive (see Bigand & Poulin-Charonnat, 2006 for a discussion). All music listeners should be considered to have received some degree of musical experience from their environment, the only distinction being whether this was formal or informal in nature.

In this section the role of similarity judgements in music perception and cognition has been considered, and research has been highlighted that has investigated it in music perception.

The process of cue abstraction was identified as an integral part of the process of judging similarity, and definitions of both cues and cue abstraction were presented. Studies that investigated the use of diverse cue sources in different musical contexts were also reported, and the assumptions underlying them were discussed. These assumptions were then considered in comparison to the ones proposed by linguistic research into the cue weighting phenomenon. This comparison highlighted a general difference in the proposed reasons underlying the perceptual inequality of cue sources – in current music-specific research contextual salience is considered the most important factor leading to a listener's use of a parameter as a cue source, in speech-specific research it is suggested that bias stemming from individual experience plays a greater part in determining the choice of an acoustic parameter as a source of cues. This study will consider both perspectives in interpreting its results.

2.2 Music as Sound

The most commonly accepted basic unit of musical identity in music theory is the motif (Lamont & Dibben, 2001; Eitan & Granot, 2009). This is a short melodic, harmonic or rhythmic idea that can maintain its identity under transformation (Drabkin, 2015). These transformations, repetitions, variations and modulations are considered the foundation of musical structure and fundamental to its sense of movement and progress over time (Deliege, 2001, p.238).

To use a linguistic analogy; where motifs are the words, phrases are the sentence, joining together a sequence of smaller motifs into a larger perceptual unit. White (1984) describes the melodic phrase as "the smallest musical unit which conveys a more or less complete musical thought" (p.34). A larger scale musical work typically comprises a melody, consisting of numerous consecutive phrases that are in turn made up of one or more motifs. As already highlighted, listener judgement of the similarity of melodic motifs and phrases has been used as the basis of a number of recent studies (see section 2.1).

Levitin (2009) maintains that music can be considered as eight perceptual components - pitch, rhythm, timbre, tempo, meter, contour, loudness, and spatial location. Whilst all eight of these parameters may be individually variable, as Levitin contends, and all eight undoubtedly play a role in defining the perceptual whole of music, they are not all accorded equal importance. Shepard (1999), for instance, argues that pitch and time are the most important dimensions of music and it is clear that this view is one shared by the musical establishment.

Many Western music educators, analysts and researchers regard pitch and subsequently, harmonic relationships, as the most important features of music (Cook, 1994, p.16). Conducting a search of the word ‘pitch’ in the Google Scholar search engine at the time of writing produces 2.84 million results, as compared to ‘rhythm’ (1.74m), and ‘timbre’ (0.14m). In the preface to their 2008 book Benward and Saker assert that a “thorough study of melody, rhythm, and texture is included. In this way, the authors hope to present “a more balanced view of the structure of music than those books that concentrate only on harmony and voice leading.” (Preface, p.x). Nevertheless, 16 of the 17 chapters of their book deal primarily with pitch and harmonic relations. Eitan and Granot (2009) also highlight the prevailing view among music theory scholars that only pitch-based and metrical parameters can define a melodic motif. Manipulations in other parameters - timbre, dynamics or tempo - constitute variations on a motif established by pitch and meter, whereas variation in pitch interval or metrical content would result in a novel motif category.

Despite the apparent primacy of pitch interval and metrical structure to music theory, however, research has frequently shown that music listeners, even those with formal training, tend to employ other parameters, ones which might be considered extra-musical or auditory, in music perception. Addessi and Caterina (2000) asked their participants to segment pieces by Milhaud, Webern and Maderna. The parameters identified in segmentation were intensity, timbre and rhythmic elements. Parameters used across participant groups were similar for both the Webern and Maderna stimuli, and only differed in response to Milhaud, where participants tended to employ structural features in segmentation. Musicians based their segmentations on the introduction of new material, whereas non-musicians based it on the conclusion of old. Listeners in Lamont and Dibben (2001) used dynamics, articulation and textures as the basis for similarity judgements relating to excerpts from Beethoven and tempo and dynamics to rate excerpts from Schoenberg, eschewing the use of what the authors called ‘deeper’ features. Between groups they observed significant differences in feature choice and descriptive abilities, but no evidence that listeners of either group had attempted to use thematic or motivic variation (i.e. interval or metric change) as the basis for classification. Participants in McAdams et al. (2004) based their judgements of similarity on tempo, rhythmic and melodic texture, pitch register, melodic contour, and articulation. They found differences between participant groups only in the ability to describe accurately what their judgements were based on, where formally trained participants showed an advantage. Eitan and Granot (2009) deliberately set up a similarity rating task on short artificial stimuli that presented manipulations in pitch intervals and metrical structure against manipulations in melodic contour and expression. The expression parameter, a compound of dynamics, texture, articulation and register, was the most frequent basis of ratings for both participant groups, but formally trained participants also used rhythm and melodic contour. Neither group showed any inclination to use pitch interval or interval change.

In summary, a body of results from current research seems to suggest an inclination amongst music listeners, irrespective of formal musical training, to process music along general auditory dimensions, such as timbral, textural and dynamic variation, rather than musical-syntactic ones such as interval and harmonic relations. The present study will consider both perspectives in interpreting its results.

2.3 Everyday Music Listening

In his book, What to Listen for in Music (originally published in 1957) Copland (2011), proposes three basic styles of music listening; the sensuous, the expressive and the sheerly musical. The sensuous listener, he argues, is the most common: "The simplest way of listening to music is to listen for the sheer pleasure of the musical sound itself… One turns on the radio while doing something else and absent-mindedly bathes in the sound." (p.10). Expressive listening, alternatively, is the deliberate attempt to seek the expressive qualities, emotional and thematic, in the music. Finally there is the sheerly musical approach, which, the author contends, requires the fostering in the listener of an understanding of musical form and structure, in order to more fully appreciate the composer's intentions. More recently, Huron (2002), in a presentation to the Society for Music Theory, proposed 21 discrete listening modes and claimed that the list was not exhaustive. This underlines another problem for music cognition research - understanding the true nature of common music listening experiences and practices.

In order to get a clearer picture of what a common music listening experience might consist of we turn to research that has used experience sampling methodologies to investigate engagement with music in everyday life. Sloboda, O'Neill and Ivaldi (2001) found that, while most of their participants reported incidences of musical experience occurring in the home, people were most likely to experience music whilst travelling, shopping or visiting entertainment venues. Only 3 of the 156 instances of reported musical experience featured music listening as the main activity. Episodes of musical experience were predominantly associated with positive mood changes. North, Hargreaves and Hargreaves (2004) reported that 50% of the incidences of music listening in their sample occurred within the home, most often with another person present, and that the reason most often given for music listening was enjoyment, but that only 10% of overall reported incidences related to music listening as the main activity. Participants in Juslin, Liljeström, Västfjäll, Barradas, and Silva (2008) were most often at home, listening to recorded or broadcast music. Only 5% of responses in their study related to music listening as the main activity. Again, in instances where the presence of music produced an emotion in the listener, these emotions tended to be positive. Of the music listening experiences reported by Greasley and Lamont (2011), the most commonly occurring was in the home, during leisure time, and listening to recorded or broadcast music. In this study only 2.3% reported listening attentively to music.

Taken as a whole the findings of the ESM studies form a relatively unequivocal picture of everyday music listening in Western culture. It occurs most frequently at home, although it is more likely while travelling or shopping. We listen to recorded music far more frequently than live. Music listening tends to be associated with positive mood change, and listening to music as the primary activity is also rare. In most cases it seems people are doing something else with music playing in the background, in line with the proposals of Copland (2011).

2.4 Methodological Concerns

Following this review of current research into musical similarity, cue abstraction and music listening it seems pertinent to discuss some methodological concerns, identified in the literature, which might be carried forward.

2.4.1 Context

Both Deliege (1996) and Lotto and Holt (2006) identify task context as a potential factor affecting the relative weighting of cues by listeners. Given the potential importance, then, of task to the use of salience or bias as the mechanism underlying cue weighting, the context in which a task is set seems critical. Since this study's stated aim is to approximate an everyday listening experience for Western participants, it is considered paramount that the experimental setting should be context-less, after the manner of Eerola et al. (2001), Lamont and Dibben (2001), McAdams et al. (2004) and Eitan and Granot (2009). Were we to impose an ecologically artificial context or goal on an experiment, such as requiring participants to analyse and describe specific aspects of their musical experience in a way that they would not naturally do, it seems likely that the results would be prejudiced by the task. In the kind of experimental setting proposed by this study, this could lead to participants highlighting cue sources as communicative of a process or device that they would not pay such attention to in a context-less everyday listening situation.

2.4.2 Control and Specificity

Another topic of relevance to this study is raised by Eitan and Granot (2009), who point out that previous research into motivic perception, including some already cited here (Eerola et al., 2001; Lamont & Dibben, 2001, McAdams et al., 2004), has frequently used complex naturalistic music as their auditory stimuli. This, in their opinion, makes the task of determining the relative contribution of musical parameters to perception harder to perform – a point conceded by some authors (Eerola et al., 2001, p.285; McAdams et al., 2004, p.232). Whilst we have thus far argued for a naturalistic experience for participants in the proposed study, it is necessary to acknowledge that if the goal of experimentation is to ascertain with any level of specificity which musical parameter a participant uses in classification, it must be by rigorously controlling other parameters so as to eliminate both redundancies and overlap between musical parameters. This rigorous control will unavoidably result in stimuli that are less like real music, but should lead to the ability to determine with some confidence the automatic listener response elicited by implicit understanding of musical materials, rather than requiring the extraction of parameter features to match to these responses.

Another consideration is highlighted in the findings of both Lamont and Dibben (2001) and McAdams et al. (2004). In both papers it is noted that non-musician participants were less able to provide consistent verbal accounts of their similarity ratings than musicians, lacking a sufficient musical vocabulary. By contrast, the same participants showed no less ability to make decisions based on the same parameters as musicians. Two points are raised by these observations, one which relates, again, to the subject of control, and another that requires separate discussion. With regard to the former point it is noted that two studies (McAdams et al., 2004; Bruderer et al., 2010) cited here collected free responses from their participants describing the reasons underlying similarity or perceived changes. This led to the necessity for the experimenters to interpret those responses into categories, which could in turn allow experimenter bias to materially affect the outcome of the experiment. It is deemed more desirable to have unequivocal categorical outcomes defined from the outset, as in the method employed by Eitan and Granot (2009), which did not allow any room for interpretation of results, either in participant responses or subsequent analyses. The latter point alluded to above is discussed in the next paragraph.

2.4.3 A Level Playing Field

Here we refer back to the observed differences between musicians and non-musicians noted in the previous section. The global precedence effect, first identified by Navon (1977), is a well-established concept in visual perception. It relates to the tendency of individuals to process a perceptual whole (the global feature) before its component parts (local features). An often used example of this is seeing a wood (global) or individual trees (local). Global precedence has been demonstrated in music cognition (Ouimet, Foster & Hyde, 2012). Most notably the global/local distinction has been presented in relation to the representation of melody as a contour, a succession of upward and downward movements in pitch (considered global), or as a series of pitch intervals (considered local) (Mottron, Peretz & Menard, 2000, p.1058-1059). Research suggests that individuals with a greater degree of formal musical training exhibit a greater ability to process both global and local features than those without (Fujioka, Trainor, Ross, Kakigi & Pantev, 2004).

The implications of these findings are twofold. Firstly, a participant with greater formal training may exhibit a greater tendency to process local features, such as pitch intervals, than a participant with only informal training, particularly in more challenging listening situations (Messerli, Pegna & Sordet, 1995, Peretz & Morais, 1987). Secondly, since the successful processing of pitch interval may depend to a greater extent on musical experience than the processing of contour features (Fujioka et al., 2004), it is submitted that the inclusion of such a feature would create an imbalance in a task in favour of one participant group from the outset. This is a situation deemed undesirable.

One of this study's main goals is to explore the relative weight of different musical parameters as cue sources to individual listeners. It is proposed, therefore, to only offer parameters for consideration that are available to all participants, and that fall under the more basic, global domain of processing (Trainor, Desjardins & Rockel, 1999). Furthermore, it is considered critical to the outcome of this study that it should not require participants to perform tasks that automatically disadvantage one group, such as the verbal description tasks highlighted in the previous section.

2.4.4 Experimental Paradigms

The final point it is necessary to raise is one of ecological validity in experimental paradigms. Much has been made in recent research of trying to capture data in more ecologically valid settings, using naturalistic stimuli, which, it is argued, allows greater generalization of results. Whilst this is a laudable goal, many fundamental problems in current approaches remain. As we have seen in section 2.3.2, the most common music-listening experiences in our culture occur in a listener's home environment, where they have some degree of control over the listening material (Sloboda et al., 2001; North et al. 2004; Juslin et al. 2008). Such experiences are associated with a generally positive mood outcome of listening to music, probably with at best diverted attention, while listeners are engaged in other activities, a finding that stands in stark contrast to the standard experimental paradigm applied in the music literature – one of fixed attention (see Deliege, 1989, p.213).

The study proposed here is exploratory in nature. In the first instance the goal is to establish explicitly whether the phenomenon of cue weighting exists in music listening, and as such no alternative presents itself but to apply a fixed attention paradigm to this first stage of the research. However, it is acknowledged that to truly present an ecologically valid setting to participants a diverted attention task would seem more analogous of most people's daily musical experiences as established by the summarised outcome of ESM studies.

2.5 Summary

In this Theoretical Background section, similarity judgements, cues, cue abstraction and cue weighting in relation to music perception and cognition were defined and discussed, with reference to previous research from both visual and auditory, musical and linguistic, domains. The assumptions – those of salience and experiential bias – that seem to underlie the cue abstraction process, according to different fields of research, were also laid out and considered. The ways in which listeners have tended to deconstruct music in other listening tasks were reviewed, and the ways in which listeners commonly experience music in their daily lives were evaluated and summarised. Finally, relevant methodological points raised by the existing literature were discussed with regard to the proposed study. In the next section the proposed experimental method will be clearly and rigorously defined, and rationale and assumptions also proposed and discussed.

3 METHOD

3.1 General Design and Assumptions

In the experiment reported here, the primary aim was to investigate to what extent listeners classifications of musical stimuli relied on a controlled set of parametric variables, and, further, to determine the task-specific relative weight of each variable to both individual participants and groups made up of those participants. Secondary aims were to set the experiment in a context-less task environment, i.e. without explicit instructions to pay attention to a given feature or device, to use simple, culturally ubiquitous musical materials, to use parameters that would be discriminable to all participants, and to present a task environment that would favour typical automatic cognitive processing. In short, to present listeners with as close an approximation of an everyday listening experience as possible given the experimental remit.

Experimental stimuli were manipulated using a 3 x 2 factorial design. Each stimulus comprised a melodic phrase with a duration of 6.5 seconds, consisting of 9 or 10 notes. These were presented to participants in eight variable states, representing all possible combinations of two different conditions in three musical parameters. All other parameters were held constant. These eight variable states formed one matrix. The experiment was divided into two parts, with four different matrices of eight stimuli presented to each participant per part.

In Part 1 (P1), four matrices were presented to participants that explored all possible combinations of melodic contour (C), meter (M), distortion (D) and loudness (L). Melodic contour and meter, which is here defined as metrical and agogic accent, were specifically chosen as more musical parameters. As discussed in section 2.2, manipulations of these parameters, according to music theory, would result in the establishment of new musical motif categories. Distortion and loudness were chosen to represent more specifically auditory parameters. Distortion is here defined as the timbral change resulting from the application of a distortion effect to the original signal. Distortion applied as a deliberate musical effect has been a frequent device in popular music since the mid-20th century (Horn, Laing, Oliver & Wicke, 2003, p.286). It has two dimensions which might make it a strong determinant as an auditory parameter. Firstly, a distorted sound presented to the ear may be perceived as louder than the same sound without distortion (Rumsey & McCormick, 2006, p.30) even when amplitude is matched. Secondly, that perceived distortion in an auditory signal can be quite divisive among listeners. Florentine, Popper & Fay (2011, p.200) point out that a loud distorted sound is usually considered annoying, but that it can be perceived as enjoyable by music listeners who have a preference for styles in which it is a feature. Consequently, distortion may be associated with negative and positive music preferences for different listeners. Distortion is frequently used in pop and rock music as a device to give choruses and refrains a higher level of activation. Well known examples of this include the Rolling Stones' "(I Can't Get No) Satisfaction" (1965) and Nirvana's "Smells Like Teen Spirit" (1991). Loudness as we define it here is the manipulation of amplitude. Research has shown that some listeners may use perceptual variation in loudness as a source of cues in both speech and music (e.g. Kochanski, Grabe, Coleman & Rosner, 2005; Luo, Masterson & Wu, 2014).

In Part 2 (P2), a different set of four parameters is used. One that is commonly manipulated musically – meter (as P1) - and three auditory – instrument timbre (T), here defined as the difference in tone quality between two instruments, distortion (as P1) and loudness (as P1). Timbre perception remains one of the most difficult topics to tackle in music research (see Hajda, Kendall, Carterette & Harshberger, 1997, for a discussion of the issues involved), yet it has been shown to play an important role in the recognition of musical materials (Poulin-Charronnat, Bigand, Lalitte, Madurell, Vieillard & McAdams, 2004), and the demarcation of musical structure (McAdams, 1999).

Table 1. shows a stimulus plan for each of the eight matrices created. Manipulations and controls applied to the stimuli will be discussed in detail later in this section.

Table 1. The order and parametric content of stimuli presented to participants over both parts of the task.

Part	Matrix Identifier	Parameters Available	Manipulations presented in stimuli (in order of presentation in Priming)							
1	1	C, M, D	Un	CMD	M	CD	D	CM	C	MD
	2	C, M, L	CML	Un	CL	M	ML	C	CM	L
	3	C, D, L	Un	CDL	L	CD	C	DL	CL	D
	4	M, D, L	MDL	Un	M	DL	L	MD	D	ML
2	5	M, I, D	Un	MID	M	ID	I	MD	D	MI
	6	M, I, L	MIL	Un	IL	M	I	ML	MI	L
	7	M, D, L	Un	MDL	L	MD	DL	M	ML	D
	8	I, D, L	IDL	Un	D	ID	L	IL	DL	I

Abbr. – Un Unmanipulated Melody, C Melodic Contour, M Meter, I Timbre, D Distortion, L Loudness

The main task in both parts of the experiment was to perform a series of classifications on the stimuli presented. Each classification required participants to group stimuli which “seemed to belong together”. The first classification (the primary grouping, T1) required participants to divide the initial eight stimuli into two groups of four. Two subsequent classifications (the secondary groupings, T2 and T3) required participants to subdivide each group of four stimuli into two further groups of two. The assumption underlying the task relates to matters discussed in section 2.1. A review of the literature suggested that listeners would either group according to parameter salience, or according to personal bias for given parameters. In either case, it was hypothesized that the primary grouping would be made based on whichever parameter was most salient or most preferred, and the secondary grouping made based on whichever parameter was the next most salient or next most preferred, in other words that groupings would have hierarchical importance.

3.2 Participants

Thirty students and researchers (fifteen male and fifteen female) from the University of Jyväskylä took part in the experiment. Participants’ ages ranged from 20 to 39 (M = 28.33). Participants were restricted to those born and brought up in Europe, North America and Australia, so as to have been exposed to predominantly Western cultural influences during developmental acquisition of music. For the purposes of this study, participants were considered as falling into one of two groups – formal and informal.

The formal group (F) (eight female, seven male, age range 23 – 39, *M* = 29.9375) had all had extensive instrumental training, ranging from 10 to 31 years (*M* = 20.33), on, in most cases, multiple instruments (range 1 – 4, *M* = 2.33), but with a focus on one particular instrument. They had all studied music as an academic subject, and had all completed a Bachelor's degree or conservatoire equivalent in a music-related topic. Finally, they had all received at least one year of explicit music theory training as part of their academic study.

The informal group (I) (eight male, seven female, age range 20 – 37, *M* = 27.13), by contrast, had had a maximum of three years of instrumental training, and in most cases (60%) had received no instrumental training at all (range 0 – 3 yrs., *M* = 0.8 yrs.). None had studied music academically, nor received any explicit music theory training.

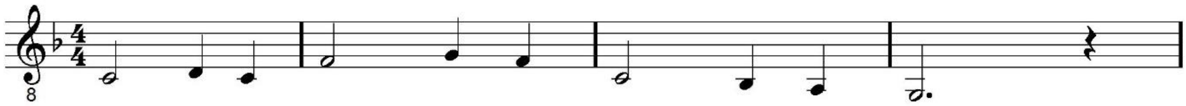
Four participants, two from the formal group and two from the informal group, reported having hearing abnormalities in the form of mild tinnitus, a condition generally associated with some degree of high-frequency hearing loss (König, Schaette, Kempter & Gross, 2006).

3.3 Experimental Materials

The materials presented to participants in this study were the eight matrices described in section 3.1. These were four matrices, comprising P1, featuring all combinations of melodic contour, meter, distortion and loudness, and presenting combinations of different states in three of these parameters in each matrix. A further four matrices, comprising P2, featured all combinations of meter, instrument timbre, distortion and loudness. The goal in all stimuli was to present two distinct states in each parameter and to minimise redundancy and overlap between parameters by applying rigorous controls.

The starting point in each part was a nine or ten note melodic phrase, composed specially for the task. These melodies were of a simple tonal style ubiquitous in Western culture, particularly in music experienced during developmental stages - lullabies, nursery rhymes, folk, religious and nationalistic songs. Musical Examples 1 and 2 show the scored, unmanipulated melodies used in P1 and P2.

Musical Example 1.



Phrase 1 - the starting (unmanipulated) melodic phrase used in P1.

Musical Example 2.

Phrase 2 – the starting (unmanipulated) melodic phrase used in P2.

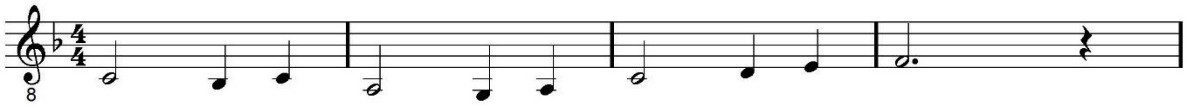
The melodic contour used in Phrase 1, the melody from P1, was quite different to that used in Phrase 2. Represented as Parsons code (Parsons, 1975) the contour of Phrase 1 is *uduuddddd and Phrase 2 is *uuuuddud. Phrases 1 and 2 were more similar metrically, differing only in the second bar. In Phrase 1 bar 2 there were three events, one of two beats followed by two of one beat. In Phrase 2 bar 2 there were two events, both of two beats, creating a different scheme of metrical and agogic accents, and one less event globally.

Stimuli used in this experiment were composed using MIDI in Cubase v5.5.3. They were then converted to audio in .wav format, at 16 bit 44.kHz, using the HALionOne virtual instrument v1.1 as a sound source. The unmanipulated timbre used in both Phrase 1 and Phrase 2 was a clarinet (Clarinet preset in HALionOne). Prior to conversion to audio the effects section of the virtual instrument was disengaged, Cutoff was set to 82, Resonance to 0, DCA Attack to 0, DCA Release to 0.05s, DCF Amount to 23% and DCF Sustain to -19.1 dB. Post-conversion, a hi-pass filter was applied to both Phrases at 55Hz with a Q of 0.9 using the Waves Renaissance Equaliser v9.3, and the peak amplitude of the audio was adjusted using normalization until the average RMS power was -15dB, as displayed by the Audio Statistics function in Cubase. In subsequent manipulations of melodic contour and rhythm the same settings and processes were applied.

3.3.1 Melodic Contour

The contour manipulation was a simple inversion of the original melody, both at the local and global level. It was applied to stimuli in P1 only. Musical Example 3 shows the melodic contour manipulation.

Musical Example 3.



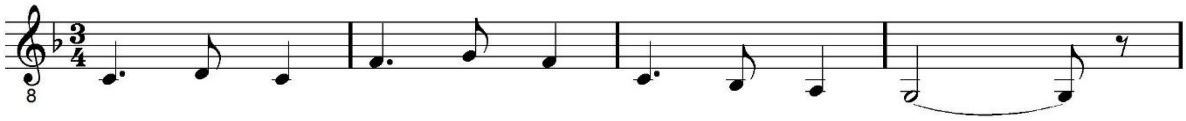
Phrase 1 after melodic contour manipulation.

Manipulation of the melody in this manner was not perceptually apparent until the onset of the 2nd event. Represented as Parsons Code the unmanipulated melody appears as *uduuddddd, the manipulated as *dudduuuuu. Intervals in the transformation were not preserved, as it was considered more important that the melody should remain consonant and in key. As such, it is acknowledged that although this manipulation is described as one of melodic contour, it was also possible that participants with sufficient expertise could have used pitch interval differences in classification. To control against this would have necessitated using melodies without conventional interval or tonality, such as those presented to participants in Eitan & Granot (2009). This was deemed counterproductive in this study, the object of which was to present participants with a simple, everyday musical experience for consideration. However, following the points discussed in sections 2.4.3 and 3.1, we proceeded on the assumption that melodic contour was the more likely means of classification across all participants, without further control measures.

3.3.2 Meter

The metric manipulation involved a change in metrical and agogic accents, so as to suggest a shift from simple duple to simple triple time. This manipulation was applied to stimuli in both P1 and P2. Musical Example 4 shows the manipulation as it applied to Phase 1.

Musical Example 4.



Phrase 1 after metric manipulation.

The duration and global event density of the manipulated phrase were preserved by altering the tempo from 140 bpm in the unmanipulated version to 107.34 bpm in the manipulated version. This also meant that the duration of the first note in each phrase was identical and that the manipulation was not perceptually apparent until the onset of the 3rd event.

3.3.3 Timbre

Instrument timbre manipulation was applied in P2 only. The alternative timbre used was a flute (Piccolo preset in HALionOne). Prior to conversion to audio the effects section of the instrument was disengaged, Cutoff was set to 100, Resonance to 0, DCA Attack to 0, DCA Release to 0.02s, DCF Amount to 4% and DCF Sustain to -21.4dB. Initially, the same equalisation and normalisation settings were applied to the flute timbre as those that had been used on the clarinet timbre.

Upon comparison of the clarinet and flute stimuli through the headphones used by participants in the experiment (see section 3.3.6) it was noted that although the stimuli were normalized to the same average RMS power the flute timbre sounded perceptually ‘nearer’ to the listener than the clarinet. A professional sound engineer was employed to make appropriate adjustments to the new timbre to counteract this. Further equalisation cuts were made to the flute timbre - a hi-pass filter was applied at 125Hz with a Q of 0.9. A bell filter was applied at 100Hz with a Q of 0.5 and a gain of -8.0, and a second bell filter was applied at 650Hz with a Q of 1.0 and gain setting of -5.0. Finally, the peak amplitude of the audio was adjusted using normalization until the average RMS power was -18.5dB, as displayed by the Audio Statistics function in Cubase. The frequency spectra of the adjusted waveforms for Phrase 2 in both clarinet and flute timbres are shown in Figure 2.

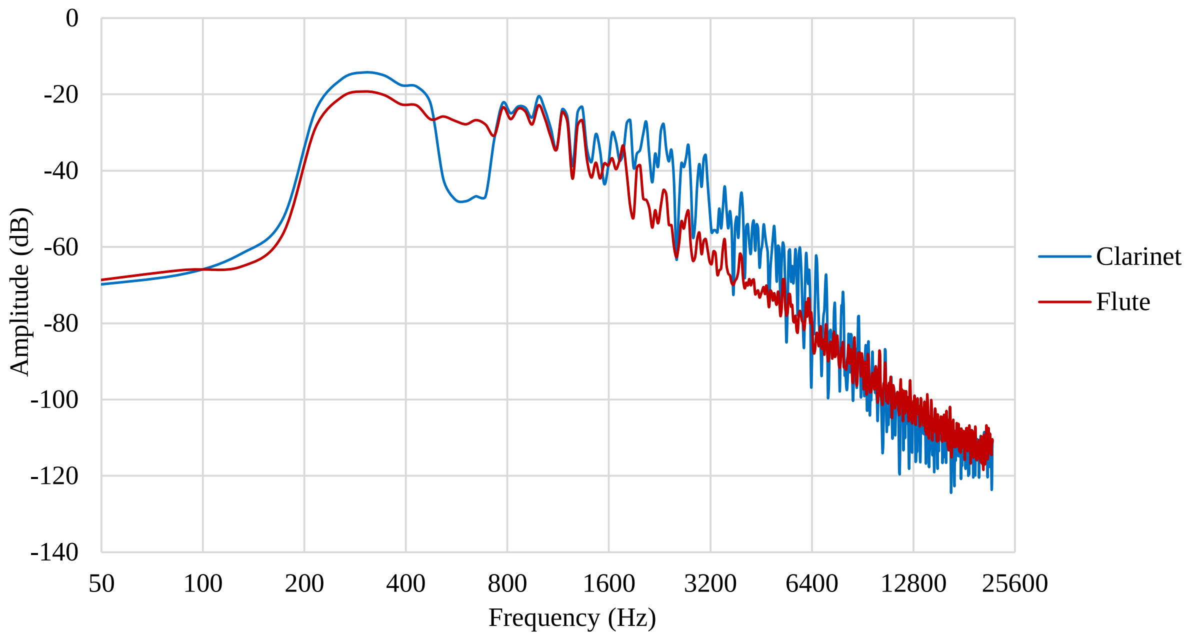


Figure 2. The frequency spectra of clarinet and flute timbres in Phrase 2.

There were three areas of audible difference between the two timbres, which are described using terminology taken from Owsinski (2013, p.64). First, the flute timbre had a mean 4.5dB less output from 129 to 517 Hz. These represent upper bass and lower mid-range frequencies. The flute had a mean 19dB more output from 517 to 689Hz, again in lower mid-range frequencies. The flute had a mean 10 dB less output from 1809Hz to 7967Hz, the upper-mid range and presence frequencies.

In contrast to manipulations in the parameters considered more musical, which in this experiment required attention to the onset of several events (three for meter, two for contour) to discern, research has shown that listeners are able to accurately detect timbral information from a tone in as little as one cycle of the sound (Robinson & Patterson, 1995). Consequently, information regarding the timbral quality of a sound was available to listeners much faster than the syntactical information of the musical features.

3.3.4 Distortion

Distortion was applied to stimuli in both P1 and P2. It was manipulated by applying the Distortion VST3 plugin to the phrases in Cubase. Boost was set to 0.5, Feedback to 5.0, Tone to 3.0, Spatial to 0.0 and Output to -5.0.

After distortion was applied, the distorted and undistorted versions were compared through the headphones used by participants in the experiment (see section 3.3.6). It was noted that at the same average RMS power the distorted timbre sounded louder than the undistorted. The peak amplitude of the distorted phrases was adjusted using normalization until the average RMS power was -16.5dB, as displayed by the Audio Statistics function in Cubase.

The frequency spectra of the distorted and undistorted clarinet in Phrase 1 are shown in Figure 3. The frequency spectra of the distorted and undistorted clarinet in Phrase 2 are shown in Figure 4, and the frequency spectra of the distorted and undistorted flute in Phrase 2 are shown in Figure 5.

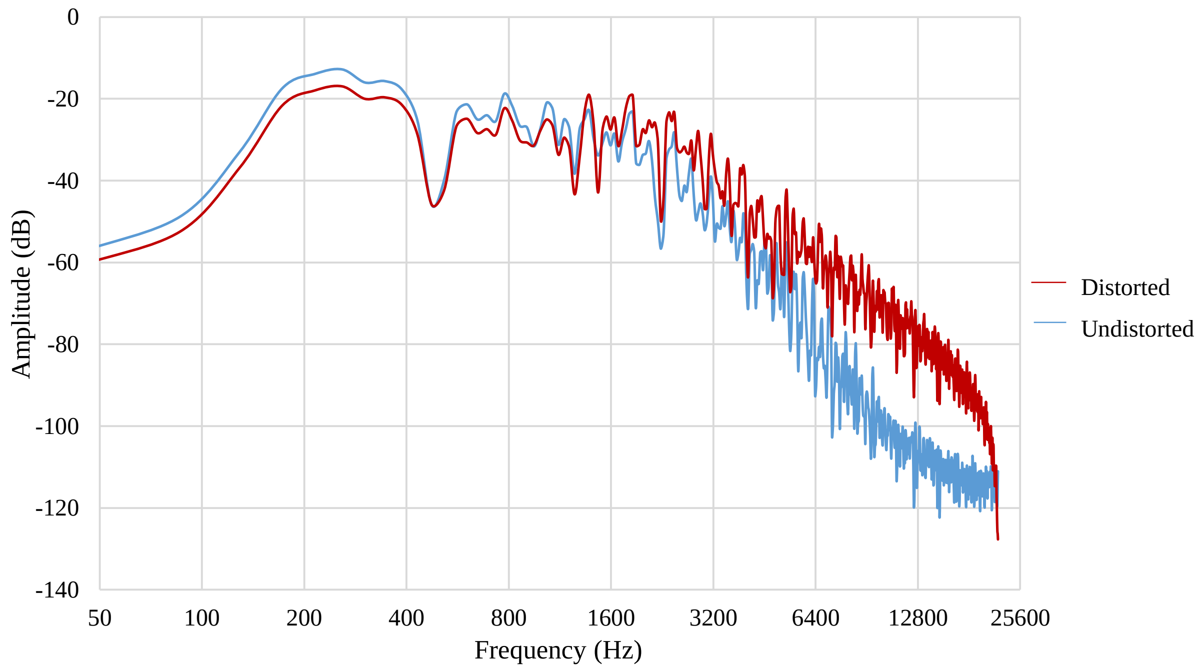


Figure 3. The frequency spectra of the distorted and undistorted clarinet in Phrase 1.

There were three audible differences between the distorted and undistorted versions of Phrase 1. The distorted phrase had a mean 3.5dB less output from 43 to 1292Hz, sub-bass to low mid-range. The distorted phrase also had a mean 8.5dB more output between 1335 and 5039 Hz, low to upper mids and presence. The distorted phrase had a mean 23dB more output from 5168 to 21619Hz, the brilliance frequencies, but this difference would likely have become inaudible due to low output from above 14000Hz.

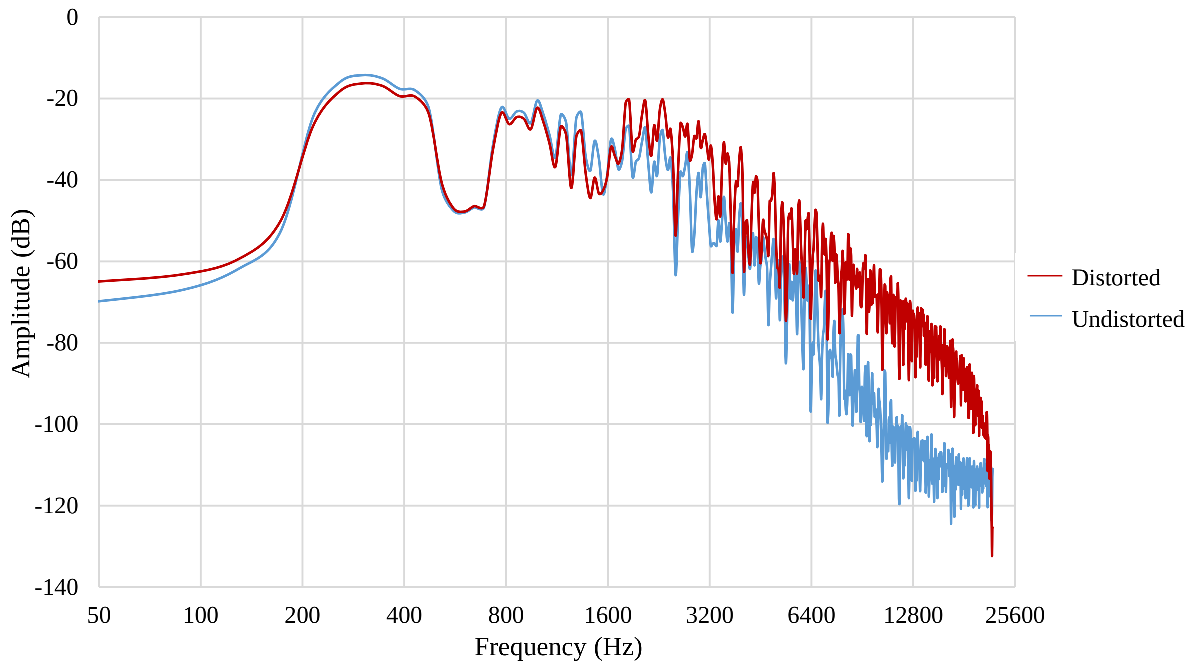


Figure 4. The frequency spectra of the distorted and undistorted clarinet in Phrase 2.

There were similar audible differences between the distorted and undistorted versions of the clarinet timbre in Phrase 2 as with Phrase 1. The distorted phrase had a mean 3dB more output from 43 to 172Hz, albeit at a very low level. Between 215 and 1680Hz the distorted phrase had a mean 2dB less output, between 1723 and 5039Hz a mean 10dB more output, and above 5168Hz a mean 23dB more output, which, again, would likely have become inaudible due to low output from around 14000Hz.

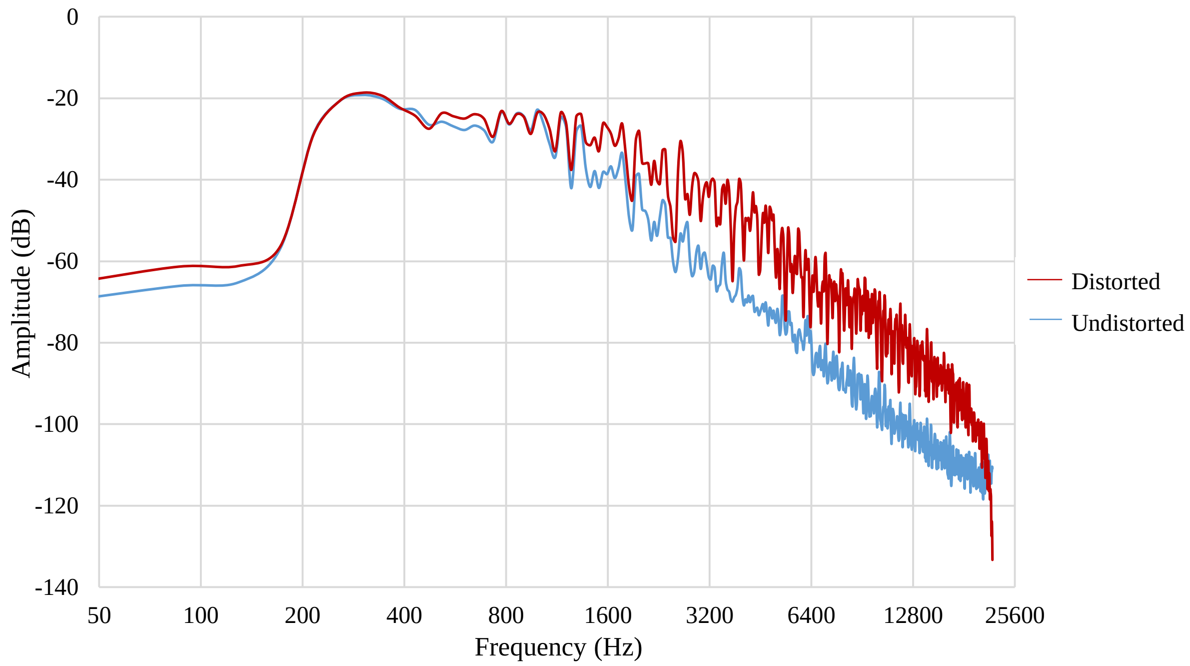


Figure 5. The frequency spectra of the distorted and undistorted flute in Phrase 2.

The distorted version of the flute timbre in Phrase 2 had a mean 3dB more output from 43 to 172Hz, again at a very low level. Between 517 and 689Hz the distorted phrase had a mean 2.5dB more output, and between 1249 and 20930Hz a mean 17dB more output, which would likely have become inaudible due to low output from around 14000Hz.

As illustrated, the distortion manipulation affected the different instrument timbres differently, tending to attenuate more low mid and bass frequencies in the clarinet timbre than the flute, while boosting a similar range of upper mid and high frequencies in both timbres. As the distortion manipulation resulted in a timbral change, the same processes applied to its detection as that of instrument timbral manipulations, meaning that it was perceptually apparent to participants very rapidly.

3.3.5 Loudness

The loudness manipulation was a reduction in overall output level. The peak amplitude of the audio was adjusted using normalization until the average RMS power was 5dB less than the unmanipulated version, as displayed by the Audio Statistics function in Cubase. This manipulation was applied to stimuli in P1 and P2 after the other procedures already described.

3.4 Setting and Apparatus

Data was gathered in a small sound-treated music studio room, selected for privacy and reasonable comfort. Experimental instruments were run in the Max 5.1 environment on a laptop computer and participants heard the stimuli over ATH-M50 closed back headphones with the volume set to a comfortable listening level.

3.5 Procedure

Upon arrival at the venue, participants were greeted and given an outline of how to proceed with the task. They were invited to access an application that presented a short set of instructions, followed by the experimental application itself. Additionally, participants were asked to complete a short questionnaire about their musical experience once they had completed the task. The experimenter then left the room so as to allow participants to relax and complete the task at their leisure.

The tone of the instructions was kept deliberately light and avoided being too directive or imperative. The matrices were referred to as grouping games that the participants would play, and the instructions briefly set out how to complete the games, including an introduction to the visual interface and mechanics of the experimental application. The main instruction to participants was to listen to the eight musical excerpts presented, then divide them into two groups of four, that "seemed to belong together", then to repeat that process dividing each group of four into two further groups of two. When participants were ready to begin they pressed the *Start* button on the experimental application.

Pressing the *Start* button on the main application triggered an initial priming playback of all eight stimuli in the matrix. The priming listen lasted 91 seconds, in which participants heard all the 6.5 second stimuli with a gap of 5.5 seconds between them, plus 0.5 seconds added to the end of the final stimulus. At the conclusion of the priming playback, participants were presented with a visual interface comprising two square boxes, one at screen left and one at screen right, with a hexagonal arrangement of icons representing the eight stimuli between them in the centre of the screen. See Appendix A for images of the main experimental application. Participants could hear the excerpts again by double-clicking on an icon to play back its associated excerpt. They then dragged and dropped the icons into the two boxes to form groups (the primary grouping, T1) and pressed a button marked *I'm Done* when the process was completed to their satisfaction. Participants were not permitted to progress to the next step unless the application detected the presence of four icons in each box. Unsuccessful attempts to progress were logged, capturing the grouping state at the time. Following successful progress, the box at screen right, with contents, would disappear and participants were instructed to repeat the grouping exercise on the four icons in the box at screen left, dividing these into two groups of two (the first secondary grouping, T2). Again, participants could not progress until the application detected the appropriate grouping, when the process was repeated with the box at screen right (the second secondary grouping, T3). At the conclusion of the second secondary grouping pressing the *I'm Done* button led the participant back to the starting screen, where pressing *Start* would begin playback of the priming for the next matrix. The entire process was repeated eight times, once for each of the eight matrices.

The order of presentation of the matrices was randomised for each participant. The priming playback for each matrix was the same for each participant. It started in every case with either the unmanipulated melody followed by the stimulus with all three parametric manipulations, or vice versa. These were alternated according to the matrix number, and subsequent stimuli were ordered so as not to unduly favour one parameter (see Table 1.). The starting position of the icons in the initial hexagonal arrangement was different but not randomised for each matrix. The icons themselves were selected to make it easier to discriminate between the excerpts visually. They used abstract objects with no prevalence of colour or form and no obvious connection between them. The allocation of each excerpt to an icon was randomised in each matrix presented to participants.

During the task various types of data were collected. The classification outcome of each grouping was recorded. From the initial press of the *Start* button each task was timed, with the timer resetting each time the *I'm Done* button was pressed. Consequently, individual timings were captured for the primary grouping (T1) which included playback time for the priming listen, and both secondary groupings (T2 and T3) for each matrix. The frequency of participant-initiated playback of the stimuli for each of the grouping stages was also captured, and the time spent listening to each individual playback of a stimulus.

At the conclusion of the grouping task participants completed a survey with demographic information and questions about their musical history.

3.6 Weighting System

In line with the assumption on which this experiment was based, that participants would group in the primary and secondary stages according to the parameters to which at each stage they afforded relative primacy, it was reasoned that the categorical outcomes of the primary and secondary groupings could not be treated as having equal weight. This would preclude the use of the frequency of parametric choice, commonly employed as the indicator of cue weight in previously identified studies (e.g. McAdams et al., 2004; Eitan & Granot, 2009; Bruderer et al., 2010; Granot & Eitan, 2011). Instead, it was considered necessary to contrive an appropriate system of weighting for each grouping outcome.

It was apparent that the larger the difference in the weighting scores attributed to the different groupings, the greater subsequent differences in weighting would appear. Consequently, a general approach was taken to keep scores as low as possible while still appropriately reflecting the hypothesized hierarchical differences between groupings. A primary grouping was awarded four points, since it had resulted from the grouping of eight stimuli into two groups of four. The initial inclination was then to award two points for each of the secondary grouping outcomes. However, if a participant had consistently chosen the same parameter in both halves of the secondary grouping, the result of a score of two in each of the secondary groupings would be a combined score of four, giving this parameter the same weight as the primary one. It was consequently decided to award a secondary grouping parameter one point for each grouping stage so that in a consistent grouping the total score would be two.

In a hypothetical situation where a participant grouped consistently throughout the task, i.e. consistently afforded the same parameters primacy, the resulting scores formed a perfect hierarchical arrangement with scores of 12 for the primary parameter, 8 for the secondary, 4 for the tertiary and 0 for the quaternary. See figure 6 for an example of a consistent classification.

	M1 (C,M,D)		M2 (C,M,L)		M3 (C,D,L)		M4 (M,D,L)	
Primary	C		C		C		M	
Secondary	M	M	M	M	D	D	D	D

Figure 6. An example of a consistent grouping for P1.

When represented as a histogram this would create a distribution that could be compared, using a chi-square goodness of fit statistic, to the null hypothesis that all parameters were weighted equally, with a score of 6 per parameter.

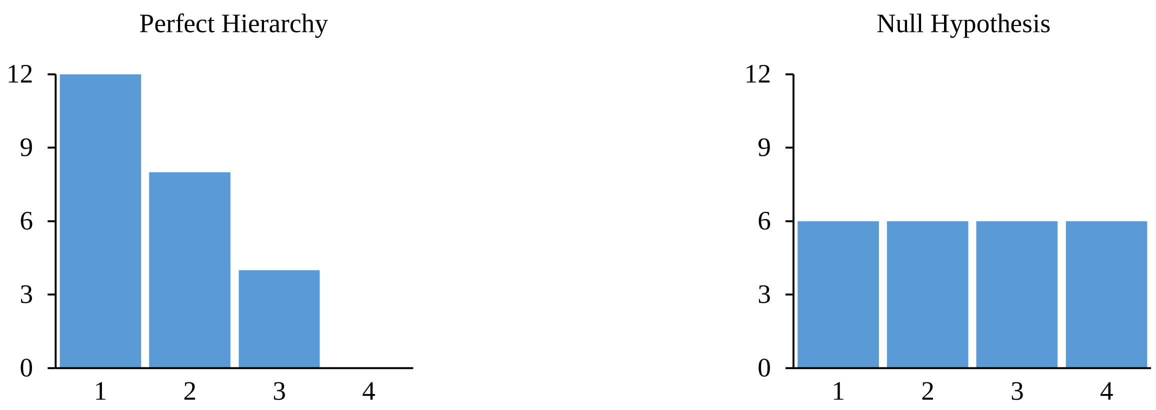


Figure 7. Hypothetical distributions illustrating consistent grouping and equal cue weight

In a situation where a participant grouped according to a parameter at every opportunity (eight primary and sixteen secondary groupings per participant) a total of 24 points was available for the construction of an individual schema of cue weight. This weighting scheme also highlighted two ways in which an individual's schema of cue weight could achieve a result of non significance – the null hypothesis, where all points were accounted for implying that all parameters were perceived successfully, but cues were weighted equally, and a second situation, where failure to perceive the parameters available resulted in a weaker schema that was also not significantly different.

3.7 Statistical Procedures

In the treatment of categorical data resulting from participants' grouping decisions, we applied two tests based on the chi-square statistic, goodness-of-fit and the test of independence. Proximity was tested using Pearson's correlation coefficient (*r*). In dealing with the task processing data, ANOVA was employed, but not in the traditional form. Due to the necessity to exclude outliers from the data to control against an order effect and other irregularities, the use of general linear models, such as the repeated-measures ANOVA was not possible, because general linear models cannot process variables with different *N*. Where there are listwise missing data points (i.e. across rows of data) the entire row is simply ignored. Consequently it was necessary to develop a fixed effects model in a linear mixed model format to process the data. The fixed effects model was based on the standard repeated-measures ANOVA, employing a compound symmetric covariance matrix, but using maximum likelihood estimation that allowed for different group sizes. An additional benefit of the linear mixed model approach is that it is not sensitive to deviations in sphericity. Consequently, sphericity is not reported in the results. For an introduction to and explanation of the use of linear mixed models see Garson (2013).

3.8 Summary

In this section a clear and rigorous account of the methods and materials employed in the implementation of the melodic similarity experiment has been given. This included definition of the assumptions underlying the research, the rationale for parameter selection, the formulation of a proposed cue weighting system, the statistical methods employed in analysis and a full description of the experimental design and procedure. In the next section the results of the experiment and the outcomes of the supporting statistical analysis are reported.

4 RESULTS

Abbreviations:					
Parameters –	Melodic Contour	C	Parts -	Part 1	P1
	Meter	M		Part 2	P2
	Timbre	I	Groupings -	Primary	T1
	Distortion	D		First Secondary	T2
	Loudness	L		Second Secondary	T3
Groups -	Formal	F	Matrices -	Matrix 1	M1
	Informal	I		Matrix 2	M2 etc.

In this section the results of the similarity experiment are reported, with supporting statistical analysis to highlight areas of importance. Results are presented with reference to the stated research questions:

- Do participants show evidence of individual cue weighting in the use of parameters available for musical similarity judgements?
- If cue weighting occurs does it relate more closely to salience or bias?
- What differences in parameter selection, task processing and performance are exhibited by participants with differing levels of formal musical training?

4.1 Individual Use of Parameters

The total number of grouping decisions made by participants in the task was 720. Of these, 690 (95.8%) grouping decisions were based on a parameter, 30 (4.2%) were not. These 30 comprised 11 primary groupings and 19 secondary groupings. Taking into account the three groupings necessary to complete the task there were 210 possible outcomes for classification. In the primary grouping, there were 35 ways to group the excerpts uniquely, since the order of groupings is not a factor (a grouping of 1234 left and 5678 right is the same outcome as a grouping of 5678 left and 1234 right). In each of the secondary groupings, there are three possible outcomes. The final figure is arrived at by multiplying 35 x 3 x 3. Of these 210 possible outcomes only 9 resulted in a parametric classification in all groupings. Participants grouped using a parameter in all groupings in 225 (93.75%) of 240 completed matrices. This was a significant result ($\chi^2(1)= 4666.5$, $p < .0001$ with Bonferroni correction), indicating that participants were not grouping according to parameters by chance.

The most frequent parameter used in primary groupings in P1 was contour, accounting for 49 (40.8%) of 120 responses. In P2, the most frequent parameter used in primary groupings was meter, accounting for 53 (44.2%) of 120 responses. The most frequent parameter used in secondary groupings in both P1 and P2 was distortion, accounting for 88 (36.7%) and 92 (38.3%) of 240 responses in each part. Table 2 shows the individual frequencies of each parameter used in each grouping in each part of the task.

Table 2. The frequency of parameters used in each grouping in each part of the task.

Part	Parameter	C	M	I	D	L
P1	Primary	49	33	x	29	5
	Secondary	36	77	x	88	31
P2	Primary	x	53	38	19	3
	Secondary	x	41	77	92	19

The proposed weighting scheme was applied to each participant's grouping decisions. The outcome of this was the creation of two schemas of parameter weighting per participant, one for P1 and one for P2. These schemas were tested against the null hypothesis that all parameters were afforded equal weight using a chi-square goodness-of-fit test that assumed equal scores across all parameters. See Appendix B for individual schemata and goodness-of-fit test results. The result of goodness-of-fit testing was that 49 (81.66%), 26 for P1 and 23 for P2, of 60 individual schemas were significantly different from the null hypothesis. Of the remaining 11, four (6.66%) were not significant due to equal weighting of the parameters. In the remaining seven, participants had lost at least five points (one primary and one secondary grouping) through failure to use an available parameter in grouping. If a schema was not significantly different from the null hypothesis parameters could not be considered to be weighted. Although some of the non-significant schema showed tendencies toward weighting of the parameters, they were excluded from subsequent analyses on individual parameter weighting.

The significant schemas comprised 13 unique hierarchical arrangements of parameters for P1 and 10 unique hierarchical arrangements for P2. This difference was not significant. Only two hierarchies, C/M/D/L[1] (eight, 26.66%) and C/D/M/L (four, 13.33%) accounted for more

[1] A slash is used here to denote a hierarchical layer. C/M/D/L indicates a hierarchy with four layers of relative weight, heaviest at the top. Where a participant weighted two parameters equally a comma is used to denote this.

than two participants in P1, not taking into account differences in the relative weight of hierarchical layers. If weighting variation was taken into account, i.e. considering hierarchies that featured the same hierarchical organization of parameters but different weight distributions (e.g. 12/8/4/0 and 10/8/6/0) as different, then only one schema, C/M/D/L with a distribution of 12/8/4/0, accounted for more than two (six, 20%) participants. Two hierarchies, M/I/D/L (ten, 33.33%) and I/D/L/M (three, 10%) accounted for more than two participants in P2, not taking into account weighting variations. If weighting variation was taken into account then only one schema, M/I/D/L with a distribution of 12/8/4/0, accounted for more than two (seven, 23.33%) participants. Table 3 shows the frequency of role for each parameter in individual significant schemas for each part of the task.

Table 3. The frequency of role for each parameter in individual schemas.

	P1				P2			
Role	C	M	D	L	M	I	D	L
Primary	14	8	6	0	16	8	2	0
Secondary	4	11	9	2	3	12	8	0
Tertiary	5	5	10	4	1	3	13	3
Quaternary	3	2	1	20	3	0	0	20

The weight distributions of the significant schemas were observed to fit into several general types. 12 unique distributions were identified in the significant schemas. These distributions were tested for proximity using the Pearson correlation coefficient. Values of .986 and above were considered indicative of strong similarity. Table 4 shows the proximity matrix with internal borders representing the division of the distributions into classes based on proximity scores, and shading highlighting values of .986 or above.

As can be seen in Table 4, the relationship between the distributions was reasonably clear except in the case of distribution 8. This distribution's score indicated greater similarity to distribution 1 than to distribution 9, but it was most similar to distribution 10, which was equally and highly similar to both 8 and 9. Distributions 8, 9 and 10 were therefore classed together. The obtained proximity scores established five types of distribution form, which are shown in figure 8. In each type an archetype was identified, which was the distribution that

C,M/D/L indicates a hierarchical arrangement with two primary parameters, a clear tertiary and a clear quaternary.

accounted for the greatest number of participant responses, rather than the distribution which was most similar in proximity to the others.

Table 4. Proximity scores for the 12 identified unique distributions from the significant schemas

Distribution	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
1. 12/8/4/0	1.000											
2. 12/8/3/1	.988	1.000										
3. 12/8/3/0	.996	.998	1.000									
4. 10/8/6/0	.956	.901	.929	1.000								
5. 12/6/4/2	.956	.963	.958	.857	1.000							
6. 12/6/3/2	.947	.969	.958	.823	.994	1.000						
7. 12/4/2/0	.933	.944	.936	.821	.997	.993	1.000					
8. 12/7/5/0	.988	.959	.972	.963	.963	.940	.944	1.000				
9. 12/6/6/0	.949	.904	.922	.945	.945	.907	.931	.986	1.000			
10.12/6/5/0	.970	.940	.952	.940	.971	.944	.959	.995	.995	1.000		
11.10/10/4/0	.949	.932	.947	.945	.819	.816	.776	.904	.833	.857	1.000	
12.10/10/3/1	.936	.944	.949	.888	.822	.837	.784	.873	.783	.823	.986	1.000

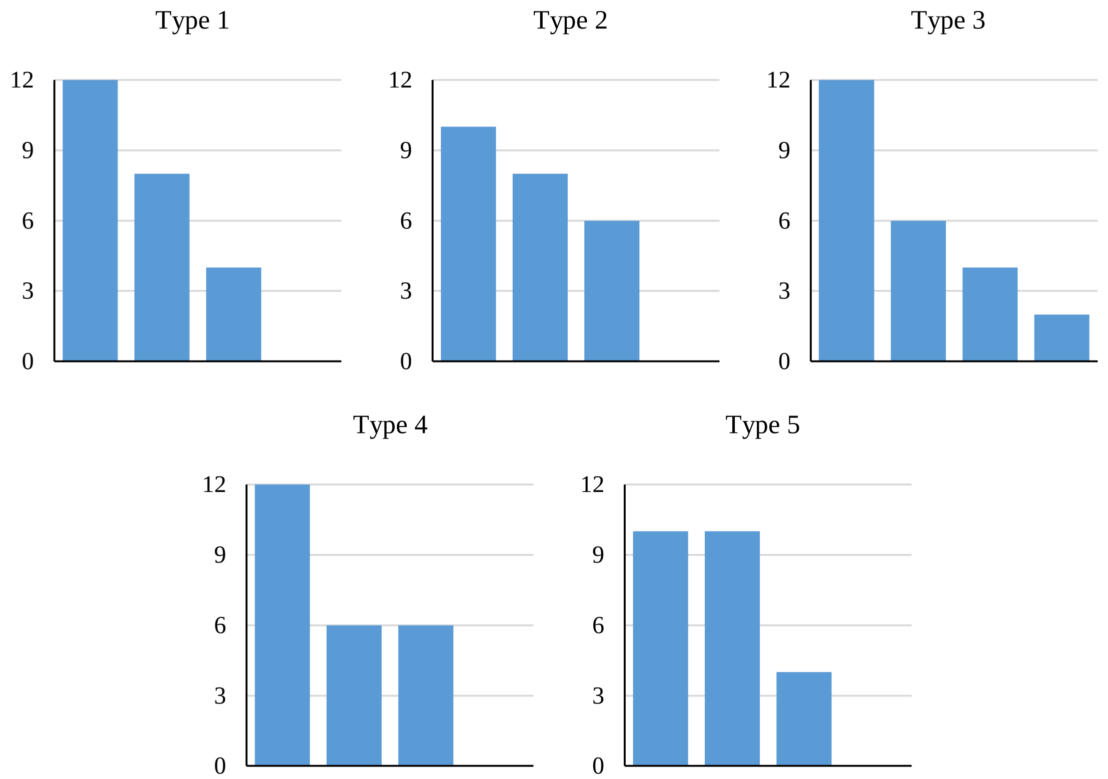


Figure 8. The five archetypes of distribution form

Type 1 was characterized by perfect hierarchical distribution. It had a significantly stronger primary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction), a significantly weaker quaternary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction for four comparisons) and a secondary and tertiary parameter that had an equal stepwise distance between them, the primary and quaternary parameters. It arose from a situation where a participant had grouped consistently, choosing the same parameters, where available, in the same roles throughout the task part. The grouping in figure 9 would result in a type 1 distribution of C/M/D/L for P1.

	M1 (C,M,D)		M2 (C,M,L)		M3 (C,D,L)		M4 (M,D,L)	
Primary	C		C		C		M	
Secondary	M	M	M	M	D	D	D	D

Figure 9. An example of a consistent grouping for P1.

Type 2 was characterized by a significantly weaker quaternary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction) and a primary three that were relatively (but not significantly) strong and relatively close together in weight, with an equal but smaller stepwise distance between the primary, secondary and tertiary. It represented a situation where a participant had grouped inconsistently, choosing a parameter considered tertiary in one matrix as the primary in two other matrices. The grouping in figure 10 would result in a type 2 distribution of C/M/D/L for P1.

	M1 (C,M,D)		M2 (C,M,L)		M3 (C,D,L)		M4 (M,D,L)	
Primary	C		M		C		M	
Secondary	D	D	C	C	D	D	D	D

Figure 10. An example of a grouping resulting in type 2 distribution for P1.

Type 3 was characterized by a significantly stronger primary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction), but a subsequent three that were relatively (but not significantly) weak and relatively close together in weight, with an equal but smaller stepwise distance between secondary, tertiary and quaternary parameters. It arose from a situation where a participant had grouped inconsistently, choosing a parameter considered tertiary in one matrix as the secondary in others. The grouping in figure 11 would result in a type 3 distribution of C/M/D/L for P1.

	M1 (C,M,D)		M2 (C,M,L)		M3 (C,D,L)		M4 (M,D,L)	
Primary	C		C		C		M	
Secondary	M	M	L	L	D	D	D	D

Figure 11. An example of a grouping resulting in type 3 distribution for P1.

Type 4 was characterized by a significantly stronger primary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction), a significantly weaker quaternary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction), and two equally weighted and proportionate (i.e. to expected proportions) secondary parameters. It represented a situation where a participant had grouped inconsistently, choosing a parameter considered tertiary in one matrix as the primary in another. The grouping in figure 12 would result in a type 4 distribution of C/M,D/L for P1.

	M1 (C,M,D)		M2 (C,M,L)		M3 (C,D,L)		M4 (M,D,L)	
Primary	C		C		C		M	
Secondary	D	D	M	M	D	D	D	D

Figure 12. An example of a grouping resulting in type 4 distribution for P1.

Type 5 was characterized by two relatively strong and equally weighted primary parameters, a relatively weak tertiary parameter and a significantly weaker quaternary parameter ($\chi^2(1) = 8$, $p < .001$, with Bonferroni correction). It arose from a situation where a participant had grouped inconsistently, in which the parameters considered primary and secondary in one matrix switched roles in another. The grouping in figure 13 would result in a type 5 distribution of C,M/D/L for P1.

	M1 (C,M,D)		M2 (C,M,L)		M3 (C,D,L)		M4 (M,D,L)	
Primary	C		M		C		M	
Secondary	M	M	C	C	D	D	D	D

Figure 13. An example of a grouping resulting in type 5 distribution for P1.

Figure 14 shows the proportional contributions of each distribution type to the significant schemas in both P1 and P2.

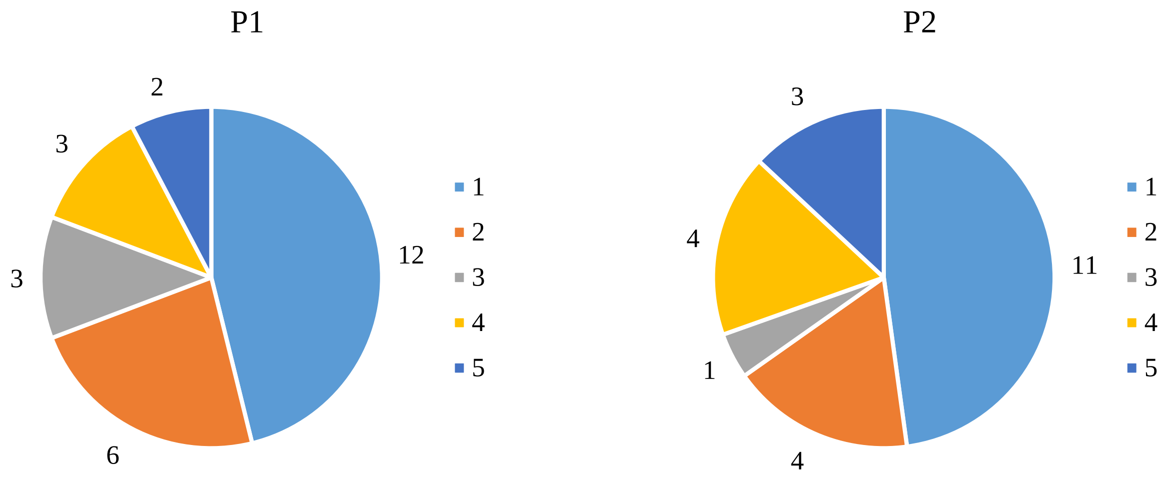


Figure 14. The proportional contributions of distribution types to significant schemas in P1 and P2

Aggregating the individual schemas, including the points from non-significant results, produced an overall schema of weighting for each part of the task:

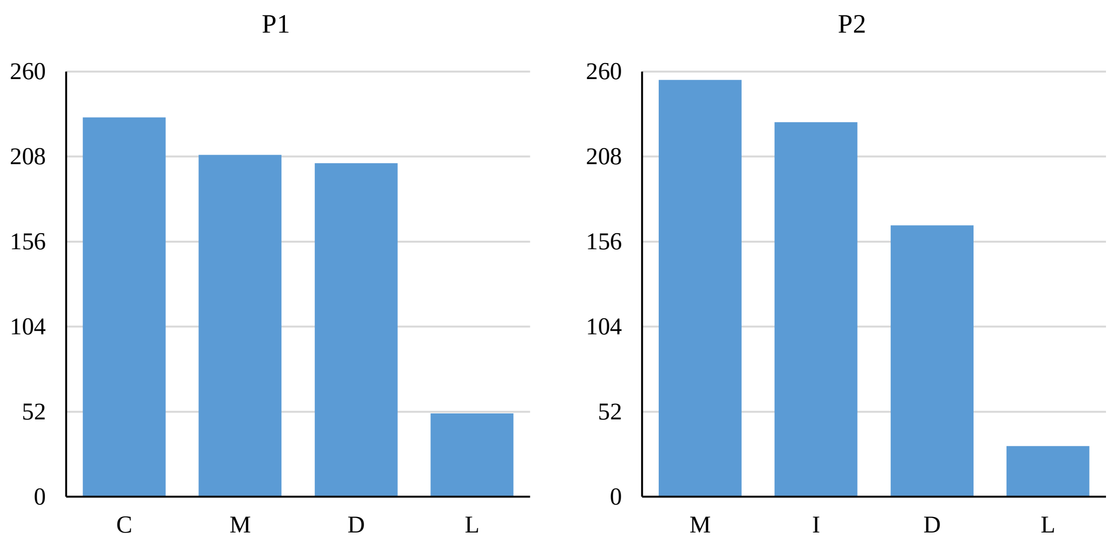


Figure 15. Overall weighting schemata for P1 and P2

The relative weighting of parameters for all participants was C/M/D/L in P1, and M/I/D/L in P2. The distribution of the schema for P1 was significantly different from the null hypothesis ($\chi^2(3) = 118.49$, $p < .0001$). In proximity it was most similar to type 2, but not at a level (.968) considered indicative of strong similarity. The distribution of the schema for P2 was also significantly different from the null hypothesis ($\chi^2(3) = 181.25$, $p < .0001$). In proximity, this distribution was most similar to type 2, at a level (.996) indicative of strong similarity.

An extension of the weighting system allowed investigation, but not confirmation, of the hypothesis that participants applied the same hierarchical weighting of parameters in both parts of the task. The scores for the parameters present in both parts were summed and the scores for each of the parameters present in only one part were doubled, proceeding on the assumption that if a parameter had been available to participants it would have been afforded a constant weight. The results of this extension suggested that 26 (86.67%) of 30 participants might be applying the same schema of parameter weight to both parts of the task. The extension method was also applied to the schemas for P1 and P2 resulting in an overall task schema, depicted in figure 16:

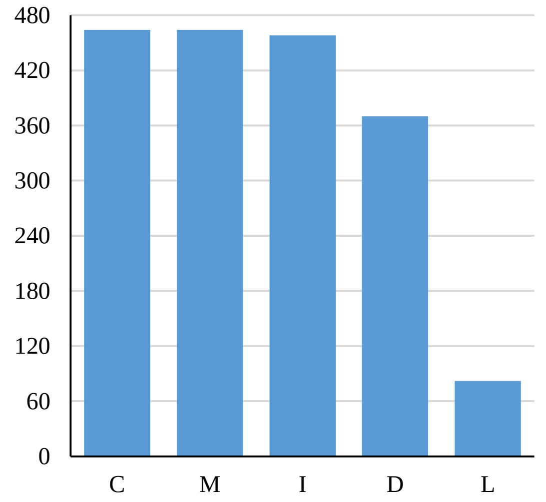


Figure 16. The overall schema, combining weighting from P1 and P2

In this hypothetical overall schema, the relative weighting of parameters was C,M/I/D/L. The distribution of the overall schema was significantly different from the null hypothesis ($\chi^2(4) = 300.44$, $p < .0001$). Proximally, this distribution was most similar to the distribution of the P1 schema at a level indicative of strong similarity (.993) and type 2, at a level (.944) not indicative of strong similarity.

4.2 Group Use of Parameters

Nine unique parametric hierarchies were used by the F group in P1 and six in P2, whereas the I group used six in both P1 and P2. These differences were not significant either between groups, within group between parts, or between groups within part. The F group did not group according to a parameter in five primary and seven secondary groupings. The I group did not group according to a parameter in six primary and twelve secondary groupings. These differences were also not significant either between groups, or between groups within groupings. Tables 5 and 6 show the frequency of role for each parameter in individual schemas attributed to each group.

Table 5. The frequency of role for each parameter in individual schemas for the F group.

	P1				P2			
Role	C	M	D	L	M	I	D	L
Primary	4	5	4	0	8	3	2	0
Secondary	3	4	5	2	2	7	1	0
Tertiary	3	3	4	2	1	1	8	0
Quaternary	3	1	0	9	0	0	0	11

Table 6. The frequency of role for each parameter in individual schemas for the I group.

	P1				P2			
Role	C	M	D	L	M	I	D	L
Primary	10	3	2	0	8	5	0	0
Secondary	1	7	4	0	1	5	7	0
Tertiary	2	2	6	2	0	2	5	3
Quaternary	0	1	1	11	3	0	0	9

Aggregating the individual weighting schemas by participant groups produced group schemas for both P1 and P2, shown in Figure 17.

The relative weighting of parameters in P1 was D/M/C/L for the F group and C/M/D/L for the I group. These were significantly different ($\chi^2(3) = 17.72$, $p < .001$) according to a chi-square test of independence. In P2 the weighting of parameters was M/I/D/L for the F group and I/M/D/L for the I group. These were not significantly different ($\chi^2(3) = 4.2$, $p < .05$). The distribution of the F group schema for P1 was significantly different from the null hypothesis ($\chi^2(3) = 47.16$, $p < .0001$). In proximity, this distribution was most similar to the I group schema for P2 (.995) and type 2 (.985) just below the level of strong similarity.

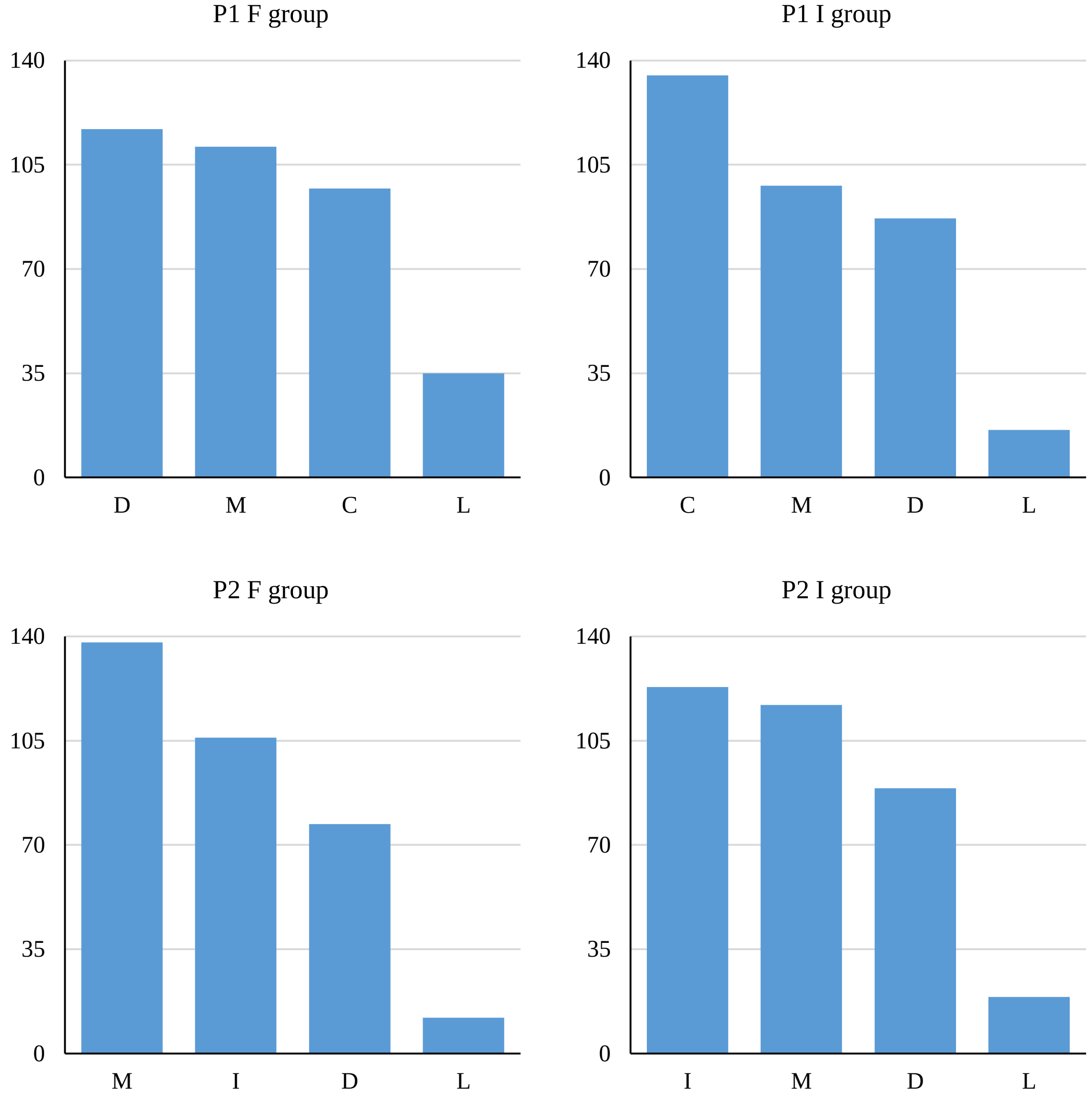


Figure 17. Group schemas for both parts of the task.

The I group schema for P1 was significantly different from the null hypothesis ($\chi^2(3) = 88.45$, $p < .0001$). This distribution was strongly similar to both type 2 (.992) and the F group distribution for P2 (.991). The F group schema for P2 was also significantly different from the null hypothesis ($\chi^2(3) = 103.67$, $p < .0001$). This schema was strongly similar to type 2 (.995) and the I group schema for P2 (.991). Finally, the distribution of the I group schema for P2 was also significantly different from the null hypothesis ($\chi^2(3) = 82.69$, $p < .0001$). It was strongly similar to both the F group P1 schema distribution (.995) and type 2 (.990).

Group schemas were also combined using the weighting extension method to explore whether the groups showed the same hypothetical application of cue weighting in both P1 and P2. In both cases, the results suggested that this was so:

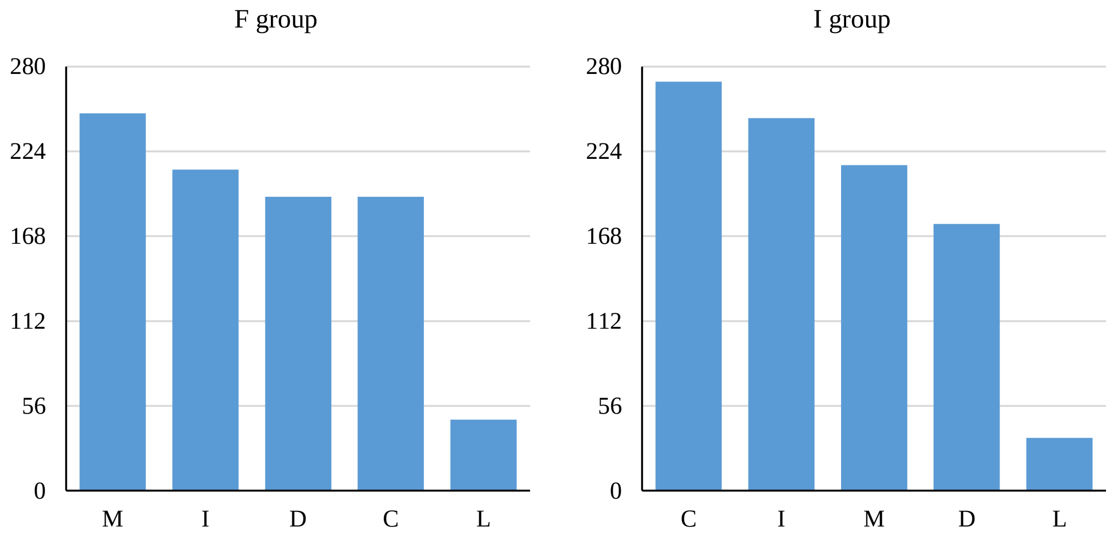


Figure 18. Group schemas for the whole task.

The relative weighting of the parameters across the whole task was M/I/D/C/L for the F group and C/I/M/D/L for the I group. Both schemas were significantly different (F group - $\chi^2(4) = 133.16$, $p < .0001$) (I group - $\chi^2(4) = 182.43$, $p < .0001$) to the null hypothesis, and were significantly different to each other ($\chi^2(4) = 18.96$, $p < .001$). The F group overall schema was most similar to type 3 distribution (.982), and the I group to type 1 (.994) and the F group's P2 schema (.993).

The final exploration that the weighting scheme facilitated was the relationship between perceptual errors in participant groups and the availability of parameters. Failure to identify a parameter in a primary grouping resulted in the award of four points of error. Failure to identify a parameter in a secondary grouping resulted in the award of two points of error. These scores were totaled for each matrix. Table 7 shows the total error per group, in each matrix and each part.

Table 7. Weighted error scores per group per matrix.

	P1				P2			
Matrix	1 (C,M,D)	2 (C,M,L)	3 (C,D,L)	4 (M,D,L)	5 (M,I,D)	6 (M,I,L)	7 (M,D,L)	8 (I,D,L)
F group	0	0	0	0	4	6	5	12
I group	6	1	5	12	1	6	0	5

There was no significant difference between the total error scores of each group. There was a significant difference between group performance in each part of the task (F Group - $\chi^2(1) =$

27, $p < .001$) (I group - $\chi^2(1) = 4$, $p < .05$). There were also significant differences between the groups' error scores in both P1 ($\chi^2(1) = 24$, $p < .0001$) and P2 ($\chi^2(1) = 5.769$, $p < .05$). In P1 the I group had 24 points of error, compared to 0 for the F group. The distribution of the I group's error scores across the matrices of P1 was significant ($\chi^2(3) = 10.33$, $p < .05$), and the error score of greatest significance was in M4 ($\chi^2(1) = 8$, $p < .01$ with Bonferroni correction). In this matrix, the I group's most heavily weighted parameter for P1 (C) was not available. In P2, the F group had 27 points of error to the I group's 12. The distribution of the F group's error scores was not significant and the score in M8 was also not significant ($p < .05$ but not low enough to satisfy the conditions of the Bonferroni correction for four comparisons). The error score in M8 was the highest for the F group in P2 and again corresponded with the matrix in which their most heavily weighted parameter for P2 (M) was not available.

Four participants who reported suffering from impaired hearing were responsible for six points of error in the task, while 26 participants without impaired hearing were responsible for 57 points. A comparison of the proportions of these errors was not significant ($\chi^2(1) = 0.79$, $p > .05$).

4.3 Task Processing

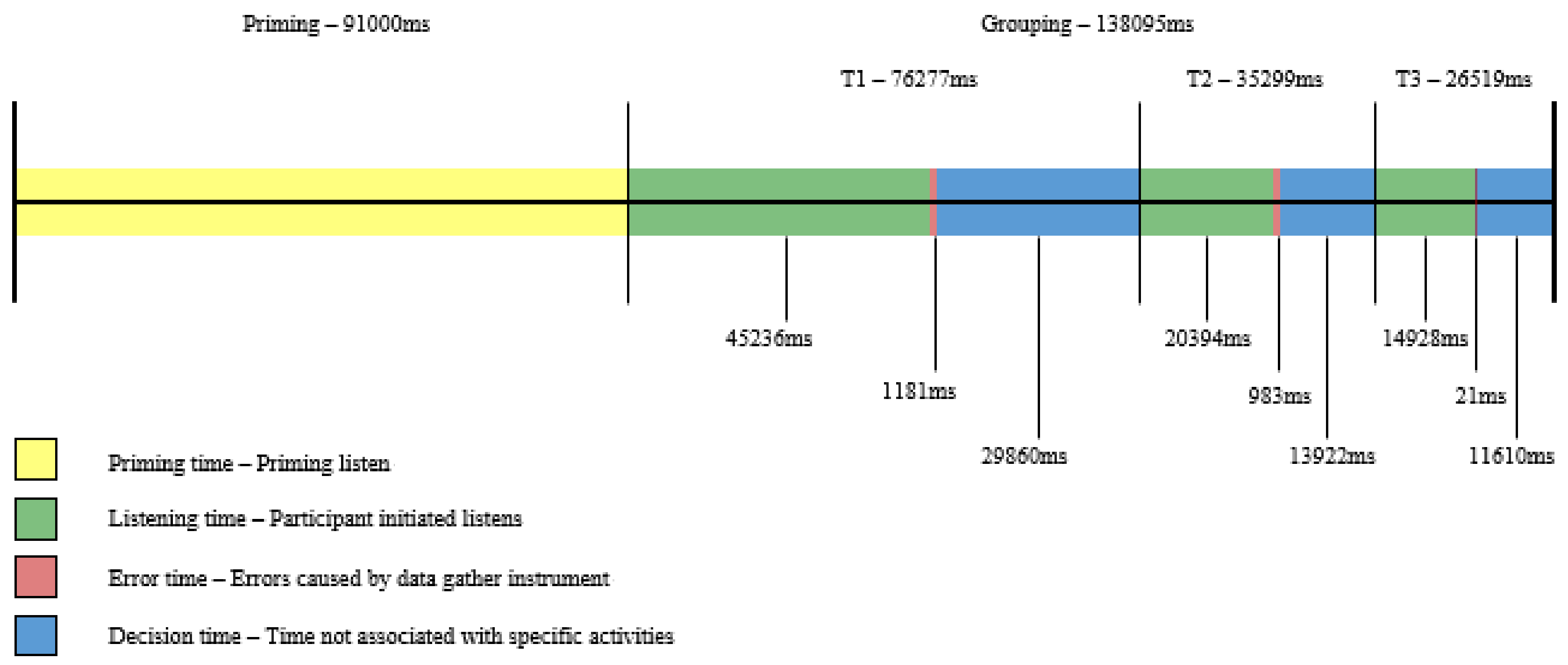
As well as the categorical outcome of grouping decisions, data was collected from participants about the duration of the task and the frequency and duration of participant-initiated listens to excerpts during task completion. Figure 19 shows the structure, in time components, of the mean completed matrix.

As described in section 3.1, participants were first exposed to the stimuli in a priming listen. During the priming listen, no grouping was possible. Task duration and frequency of listen data was captured in three chunks or tests - from the start of the priming listen to the completion of the primary grouping (T1), for the duration of the first secondary grouping (T2), and for the duration of the second secondary grouping (T3). The timer would stop after completion of T2 until the start of the playback of the priming listen in the next matrix to allow for pauses in the data gather if a participant required it. This duration was not captured. No participant was observed leaving the room or pausing between matrices. All duration data was captured in ms, to an accuracy of 5ms. T1 durations, across all participants and all matrices and after subtraction of the priming listen time, ranged from 21652ms to 241524ms, *M*=76276, *SD*=42601. T2 durations ranged from 8710ms to 211860ms, *M*=35299, *SD*=23971. T3 durations ranged from 6745ms to 116765ms, *M*=26518, *SD*=15704. T1 frequencies ranged from 5 to 76, *M*=16.63, *SD*=10.67. T2 frequencies ranged from 0 to 33, *M*=7.14, *SD*=4.80. T3 frequencies ranged from 2 to 24, *M*=5.87, *SD*=3.19. One way repeated measures ANOVAs revealed significant effects of test (T1, T2 and T3) on duration, $F(2, 690) = 233.28$, $p < .001$, and frequency, $F(2, 690) = 209.05$, $p < .001$.

Listening time in each of the three tests was calculated by summing the individual participant-initiated listen durations gathered during that test. Error time occurred as the result of the experimental instrument failing to detect the grouping status required to progress to the next grouping, and was judged as the time from when a participant attempted to progress without the required grouping to when they successfully progressed to the next grouping. Decision time was time left over not accounted for by listening or error, calculated by subtracting these from test duration. It was possible to determine whether a participant had skipped the priming listen by subtracting the priming listen time (91000 ms) from the T1 duration. If this resulted in a negative figure then priming had been skipped.

Total duration for the completion of each matrix was calculated by summing the T1 (less priming time), T2 and T3 duration values. Total frequency of listens was calculated by summing the T1, T2 and T3 frequency values. Total duration and total frequency values for each matrix for all participants and for each participant group were subjected to EDA using boxplots in SPSS. Outliers were detected from participants who had skipped priming (six matrices from one F group participant), participants who had failed to detect parameters in a matrix, resulting in longer duration and/or more listens (seven matrices, four from F and three from I), and participants who had indicated suffering from hearing impairment whose durations were longer and/or needed more listens (two matrices, one from F and one from I). In each case both duration and frequency data was excluded from subsequent analyses.

Figure 19. The mean completed matrix for all participants in all matrices.



Following these exclusions, a number of outliers remained in both variables that were noted to frequently coincide with the first matrix completed by participants. One way repeated measures ANOVAs revealed significant effects of order of presentation on total duration, $F(7, 196.12) = 19.78$, $p < .001$, and total frequency, $F(7, 192.93) = 3.55$, $p < .01$. Pairwise comparisons with Bonferroni correction revealed that the first matrix' mean duration was significantly longer than all other matrices, the second matrix' duration was both significantly shorter than the first and significantly longer than the fourth, sixth, seventh and eighth, and the third matrix' duration was significantly shorter than the first and significantly longer than the eighth. Pairwise comparisons, with Bonferroni correction revealed that participants listened to excerpts significantly more times in both the first and second matrices than in the fourth. Based on these results matrices 1, 2 and 3 were labelled the learning phase of the task, and 4, 5, 6, 7 and 8 the normal phase. Figure 20 shows plots of total duration and total frequency means and the separation of the task phases. Since this first analysis highlighted significant differences between the phases on both variables, learning phase data was excluded from subsequent analyses so as to reduce variance in the sample and the possibility of Type I error.

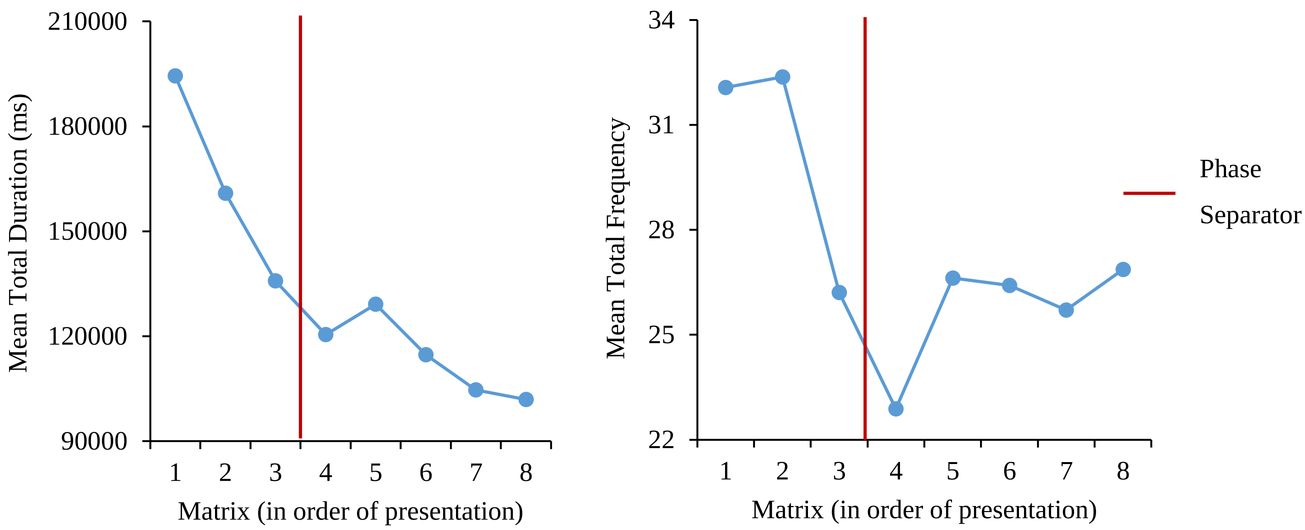


Figure 20. Plots of mean total duration and frequency for matrices in order of presentation.

A factorial repeated measures ANOVA revealed significant main effects for test (T1, T2 and T3), $F(2, 396.62) = 156.92$, $p < .001$, and group (F and I), $F(1, 30.35) = 10.98$, $p < .01$, but not for part (P1 and P2), on duration. No significant interaction of test, part and group was observed, but pairwise comparisons with Bonferroni correction revealed significant differences between the groups in P2 T1, $F(1, 147.14) = 10.02$, $p < .01$, and P2 T2, $F(1,$

147.14) = 8.59, $p < .01$. Figure 21 shows plots for mean test duration by group in both parts, with significant differences highlighted.

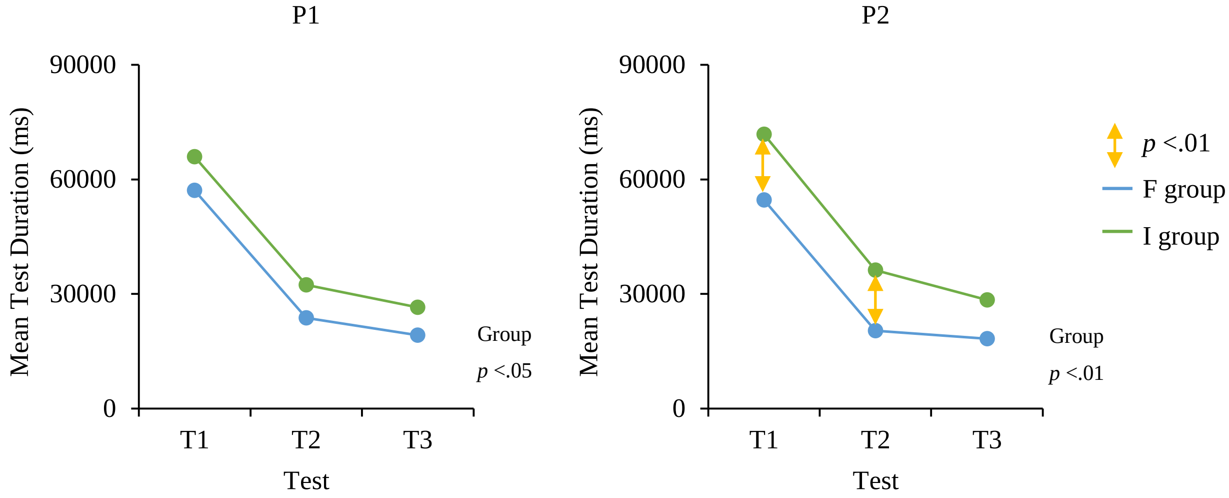


Figure 21. Plots of mean test duration by group in P1 and P2.

A second factorial repeated measures ANOVA revealed significant main effects for test, $F(2, 391.47) = 152.16$, $p < .001$, but not for group or part, on frequency. No significant interactions of any of the factors were observed, and no significant differences were revealed in pairwise comparisons.

A third factorial repeated measures ANOVA revealed significant main effects for component (listening, error and decision time), $F(2, 396.47) = 392.55$, $p < .001$, and group, $F(1, 30.15) = 15.95$, $p < .001$, but not for part (P1 and P2), on component time. No significant interaction of component, part and group was observed, but pairwise comparisons revealed significant differences between F and I groups on listening time in P1, $F(1, 210.93) = 25.43$, $p < .001$, and P2, $F(1, 164.62) = 69.39$, $p < .001$. Figure 22 shows plots for mean component time by group in both parts, with significant differences highlighted.

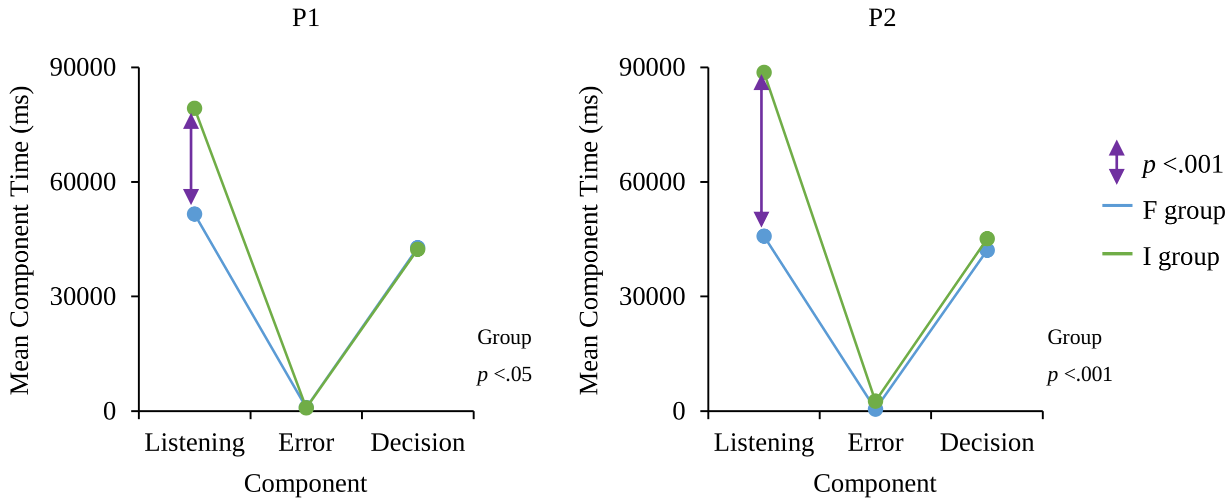


Figure 22. Plots of mean component time by group in P1 and P2.

The mean excerpt listening time for each test across all matrices for each participant was calculated by summing the listening time components for each test from all matrices, and dividing this by the sum of all listening frequencies for each test from all matrices.

A final factorial repeated measures ANOVA revealed significant main effects for part, $F(2, 145.37) = 4.289$, $p < .05$, and group, $F(1, 29.25) = 17.66$, $p < .001$, but not for test, on mean excerpt listening time. No significant interaction of part and group was observed, but pairwise comparisons revealed significant differences between F group and I group on mean excerpt listening time in P1, $F(1, 39.15) = 13.25$, $p < .01$, and P2, $F(1, 37.23) = 17.79$, $p < .001$. Mean excerpt listening times and standard deviations for the groups in P1 and P2 are presented in Table 8.

Table 8. Mean excerpt listening times and standard deviations for each group in P1 and P2.

	P1		P2	
Group	M	SD	M	SD
F	2099.86	880.76	1665.53	709.24
I	3313.29	1608.88	3243.31	1585.86

No significant interaction of test, part and group was observed, but pairwise comparisons revealed significant differences between the groups on mean excerpt listening time in all tests in both parts. Figure 23 shows plots for mean excerpt listening time by group in both parts, with significant differences highlighted.

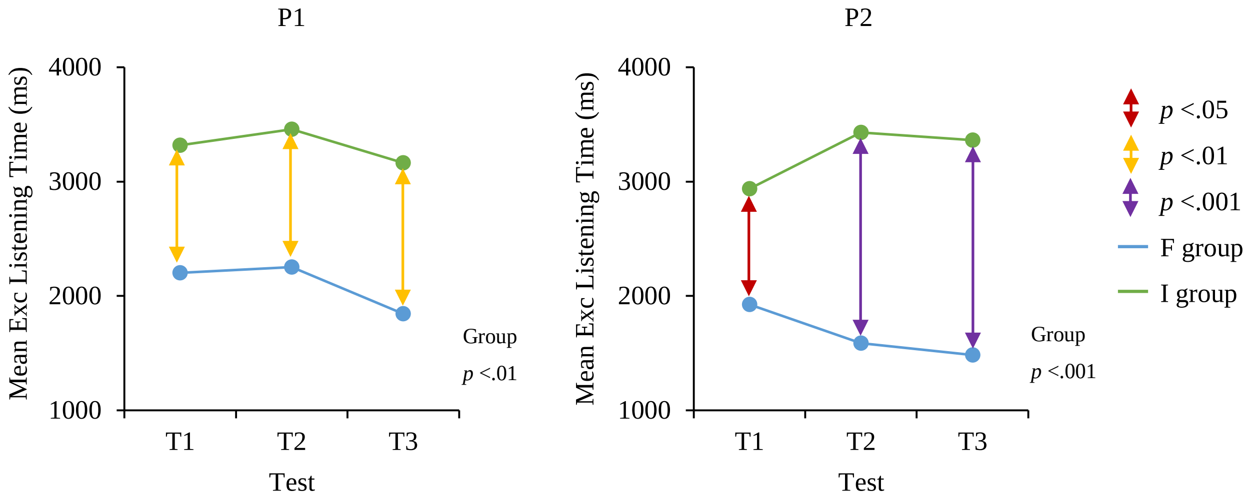


Figure 23. Plots of mean excerpt listening time in each test by group, in learning and normal phase.

A further series of factorial repeated measures ANOVAs was performed to investigate the effects of matrix (according to parameter content not order of presentation) and group on mean total duration, mean total frequency and mean excerpt listening time.

A significant main effect was observed for group on mean total duration, $F(1, 31.737) = 10.427$, $p < .01$. There was no effect of part or matrix. No significant interaction of group and matrix was observed, but between-group pairwise comparisons with Bonferroni correction revealed significant differences in mean total duration for M6, $F(1, 120.11) = 7.88$, $p < .01$, M7, $F(1, 121.82) = 12.01$, $p < .01$, and M8, $F(1, 123.72) = 8.437$, $p < .01$. Within-group pairwise comparisons with Bonferroni correction were significant for the I group, $F(7, 2.348) = 2.35$, $p < .05$, not the F, with their mean M5 duration significantly shorter than their mean M8 duration.

A significant main effect of matrix, $F(7, 117.93) = 5.08$, $p < .001$, was observed on mean total frequency, but not group. No significant interaction of group and matrix was observed, and between-group pairwise comparisons revealed no significant differences. Within-group pairwise comparisons with Bonferroni correction were significant for the F group, $F(7, 117.53) = 4.27$, $p < .001$, but not the I. Their mean M8 frequency of listens was significantly higher than that of their M3, M6 and M7.

Finally, significant main effects were observed for matrix, $F(7, 114.6) = 5.05$, $p < .001$, and group, $F(1, 29.74) = 16.37$, $p < .001$, on mean excerpt listening time. No significant interaction of part and matrix was observed, but between-group pairwise comparisons with

Bonferroni correction revealed significant differences between all matrices except M4. Within-group pairwise comparisons were significant for both F, $F(7, 114.76) = 3.01$, $p < .01$, and I group, $F(7, 113.61) = 4.1$, $p < .001$. The F group's M2 mean excerpt listening time was significantly longer than their M5 or M8. The I group's M5 mean excerpt listening time was significantly shorter than their M1, M2, M6, M7 or M8.

4.4 Summary

In this section the results of the experimental process and statistical analysis of those results have been presented. In section 4.1 the variety of individual schemas of parameter weight was documented and differences explored in parameter use and distributions. Five types of schema distribution were identified and the underlying reasons for each explored. In section 4.2 individual schemas were combined into group schemas, and the differences in parameter choice and distribution were highlighted. Differences in perceptual error rates between and within groups were found, and those errors related to the available parameters in different task parts. In section 4.3 between-group differences in task processing times were found throughout. The greatest difference was found in the mean excerpt listening times between groups, with the informal group taking significantly longer listens on average than the formal group. These results are discussed in the following section.

5 DISCUSSION

Abbreviations:					
Parameters –	Melodic Contour	C	Parts -	Part 1	P1
	Meter	M		Part 2	P2
	Timbre	I	Groupings -	Primary	T1
	Distortion	D		First Secondary	T2
	Loudness	L		Second Secondary	T3
Groups -	Formal	F	Matrices -	Matrix 1	M1
	Informal	I		Matrix 2	M2 etc.

In this section important results of the classification experiment are highlighted, both from the grouping and processing data, and discussed with regard to both this study's three research questions and the findings of other research. The research questions asked were:

- Do participants show evidence of individual cue weighting in the use of parameters available for musical similarity judgements?
- If cue weighting occurs does it relate more closely to salience or bias?
- What differences in parameter selection, task processing and performance are exhibited by participants with differing levels of formal musical training?

5.1 The Role of Advantage and Disadvantage

The first research question posed by this study sought to determine whether participants showed evidence of individual cue weighting in the use of parameters available for musical similarity judgements. The answer is a relatively straightforward and unequivocal yes. The results of the melodic classification experiment suggested that participants used individual schemas of cue weighting in parameter choices, and the diversity of those schemas was quite stunning given the small variety of simplistic and relatively monotonous stimuli employed. However, that finding only skims the surface of what this study revealed, and there were clearly a number of other factors that affected this phenomenon. These other factors also provide answers to the second research question, investigating the style of the cue weighting, and in part to the third question, regarding group differences.

The first important factor was the finding that participant groups were diametrically advantaged/disadvantaged in different task parts. The significant differences in between and within-group weighted error scores indicated an advantage for the F group and/or a disadvantage for the I group in P1, where two musical parameters and two auditory parameters were available as cue sources, and an advantage for the I group or disadvantage for the F group in P2, where only one musical parameter and three auditory were available.

This finding is unprecedented, at least in the literature that has been reviewed here. The typical results presented by other research are either that, context-dependent, participant groups show no difference in performance (e.g. Deliege, 1989; Lamont & Dibben, 2001; Bigand & Poulin-Charonnat, 2006) or that formally trained participants have a constant advantage (e.g. Fujioka et al., 2004; Stoesz, Jakobson, Kilgour & Lewycky, 2007; Schellenberg & Moreno 2010). Only one study, Carey et al. (2015), considered more than one context in which formal musical expertise could affect performance, and found that the benefits of that formal training largely occurred in situations that directly related to it. For instance, professional violinists had a greater ability to detect small changes in pitch than either pianists or non-musicians. The authors found no evidence that formal training enhanced participants' abilities in general auditory scene analysis. Whilst P2 of our task presented participants with fewer musical parameters upon which to make groupings, timbral changes are not uncommon in music of all styles and genres, and the fact remains that the parametric manipulations were still presented in a musical context not a general auditory one, which might be expected to favor formally trained participants. What effect the simplicity of this musical context, i.e. presenting manipulations in a monophonic as opposed to a homophonic or polyphonic environment, might have on listener perception is unclear, but the role of timbral change as interference in the recognition of pitch has been documented (e.g. Pitt, 1994; Poulin-Charonnat et al., 2004).

Several differences between the current study and others highlighted here might go some way to explaining this finding. First, this task was deliberately placed in a context-less setting, as discussed in section 2.4.1. Not setting out with the intention of examining a specific attribute of music-listening behavior may have afforded the opportunity to observe more general behavioral traits than would otherwise have been possible. Secondly, rigorous controls were applied in the stimuli in order to be able to say with specificity what parameter participants used to make groupings. Consequently, the possible outcomes are limited and concise –

participants have to group according to one parameter or another. This in turn allows the attribution of participant error, or failure to detect a parameter. This method is quite different from the approaches employed in many other studies, which determine advantage either through scale of reaction (e.g. Fujioka et al., 2004; Mikutta, Maissen, Altorfer, Strik & Koenig, 2014) or by frequency and consensus of response (e.g. McAdams et al., 2004; Bruderer et al., 2010). The possibility exists, of course, that the controls employed in these stimuli were not rigorous enough, and that participants were grouping according to an uncontrolled or undetected variable. However, given that the vast majority of groupings followed expected outcomes, this seems unlikely.

The second important factor was seen in the relative weighting distributions of the group schemas in P1 and P2, considered in light of the advantage/disadvantage findings. There was strong similarity between the schema distributions where groups were advantaged (F group P1, I group P2, .995) and where they were disadvantaged (F group P2, I group P1, .991), suggesting commonalities about group responses to situations of varying difficulty. Where a group was advantaged, there was less consensus and greater diversity on which parameter was the most heavily weighted. This was particularly in evidence in the F group's P1 responses, where there was the greatest diversity of individual responses of any group in any part, but this phenomenon could be seen in the contribution of individual schemas to group schemas for both groups (Tables 5 and 6 and Figure 16). Where a group was disadvantaged, however, there was a much greater consensus on a particular schema of cue weight (see Tables 5 and 6 and Figure 16 again), and much greater reliance on a single primary parameter, resulting in the greatest amount of perceptual error for each group in the matrix where that parameter was absent. In the relevant area of disadvantage, the group schema, with its greater weighting of a particular parameter, could thus be used to predict where the greatest perceptual error would occur.

With regard to the second research question, on what style of cue weighting, salience or bias, was in evidence, the answer seems to be that both occurred, but in different situations. In the advantaged state, individuals within a group showed greater variation and idiosyncrasy of response, the formal group more so than the informal, but both showing similar behavior. This suggests that, in a comfortable listening situation, an individual may exercise greater individual bias in parameter weighting. In the disadvantaged state, by contrast, there was far less variation and greater group consensus on a particular schema of parameter weighting, suggesting that participants were using perceived salience to a greater degree. Perceived salience, then, may be the fallback position in situations of perceptual difficulty. Since this approach resulted uniformly in greater perceptual error we can also conclude that individual bias tended to work better as a cue weighting mechanism. This finding could be seen to be analogous to that of Juslin (2000), when he found that although his performers used different cues to convey emotions in their performance, they were equally successful in communicating the desired emotion to their listeners.

Critically, it should also be noted that what the participant groups considered the most salient parameters in areas of disadvantage seemed to differ. The areas of group disadvantage were in different task parts, which does not readily facilitate direct comparison. However, it was clear that the formal group favored metrical and agogic accent of the available musical parameters in P1 and then focused heavily on it during their disadvantage, while the informal group used melodic contour. It seems, then, that experiential bias may also play a role in determining the parameter considered most salient in more difficult situations.

When looking at situations of advantage, by contrast, relationships were much less clear cut. It was clear that, while the formal group performed significantly better in P1, where melodic contour was available as a classification criterion, they actually used melodic contour to classify relatively little – it accounted for 26.9% of the F group weighting in P1, and for a surprisingly high number of tertiary and quaternary parameters in individual schemas, second only to loudness. Similarly, the unavailability of melodic contour and availability of instrument timbre as criteria in P2 seemed to benefit the informal group. Instrument timbre was only the most heavily weighted parameter in their group schema, however, because it was the secondary parameter of choice in more individual schemas than metrical accent, which was selected more often as the primary parameter. Furthermore, the informal group's best perceptual performance in P2 was in M7, where instrument timbre was not available as a criterion.

To summarize, it seems that no other study has found a quantifiable advantage for informal participants in comparison to formal participants in a musical situation, but, equally, no other study has considered the role of individual bias for musical parameters in such a situation. The one may conceivably lead to the other. In any case, these findings certainly warrant further investigation and could be of great significance to music cognition research going forward.

5.2 Use of Parameters

Another interesting finding of this study was that the weighting of the parameters between the groups was so different, in both hierarchical ordering and distribution. The most surprising observation was the overall reliance of the formal group on metrical accent, not melodic contour, and the much greater role of melodic contour in the informal group schema. This seemed counter-intuitive given music theory's heavy emphasis on pitch related parameters. Nevertheless, Eitan & Granot's (2009) formal participants also used metrical and agogic accent as the main classification criterion in their second experiment, preferring it over pitch-based criteria, so this finding is not unprecedented.

The role of instrument timbre as the secondary parameter in both groups' overall schemas was not nearly as unexpected, and has parallels in the findings of many studies (Deliege, 1987; McAdams et al., 2004; Bruderer et al., 2010) on the perception of musical structure.

Two further parametric weightings were worthy of note and discussion. The first was the unexpected role of distortion as the primary parameter in the F group's P1 schema, particularly when compared to the same group's P2 schema, where it was the tertiary, lesser in importance than instrument timbre. Furthermore there was no parallel in the I group schemas, where distortion played a consistently minor role. This finding is supported by those of Carey et al. (2015), who suggested that formally trained listeners have greater sensitivity to psychophysical properties of music. It also raises questions about the reaction of formal listeners to distortion in musical settings, particularly in the advantage situation which seemed to have afforded more opportunity for the expression of individual bias. This finding suggests that valence may be playing a part in individual grouping decisions as proposed in section 3.1. McAdams et al. (2004) also found evidence of affective responses as a basis for similarity judgements in their study (p.232). Further research needs to be conducted to determine the role of valenced and affective responses in similar contexts. Also of interest here is that the distorted stimuli, as discussed in section 3.1, present an illusion of loudness that seems to have been more important in group schemas than actual loudness, leading to the final point.

The role of loudness as the quinary parameter was perhaps not surprising in a musical context, despite its relative importance as a spatial location cue in more general auditory perception (Handel, 1989, p.102). What is more surprising is the overall low score attributed to loudness in comparison to the other parameters. This may be as a consequence of the employed weighting system, which in the case of a consistent grouping awarded zero points to the quaternary parameter. It may also, however, suggest that the change in this parameter was not of equal perceptual magnitude to the changes in the other parameters. The stimuli employed in this study were not subjected to any perceptual ratings prior to use, so this is impossible to determine, but suggests an improvement that could be implemented in further research.

5.3 Analytic or Holistic Listening?

The main finding evident from the processing data gathered during the experiment was that participants with formal musical training had a distinct processing advantage throughout the task, in terms of their significantly faster matrix completion times. This was, in itself, not unexpected. Other studies have found that musicians exhibit such processing advantages over non-musicians (e.g. Koelsch, Schmidt & Kansok, 2002; Amer, Kalender, Hasher, Trehub & Wong, 2013). However, this result became more informative when it was decomposed to reveal that the main difference between groups was in the listening time component, and specifically in the mean excerpt listening time per matrix.

Considering this finding from a more practical perspective, the mean excerpt listening times can be thought of as the number of events required to detect manipulations. In P1, the mean excerpt listening time for the F group was the equivalent of three full notes plus the onset of the fourth, and for the I group, the equivalent of five full notes and more than half of the sixth. In P2, as shown in Table 30, the difference became even greater. The F group's time dropped to the equivalent of three full notes, but the I group's remained practically the same. As highlighted in the various stimulus descriptions of section 3.3, if a participant had an explicit understanding of the nature of the manipulations that occurred in the stimuli then they would need to listen to the stimulus for no more than three notes. During those three notes, all possible manipulations were revealed to a listener. This level of understanding was clearly suggested by the mean listening times of the formal group, but not those of the informal group.

Why did the I group need to listen to the stimuli for so much longer than the F group? An indication was provided by the final decomposition of the listening time data into the three separate tests completed for each matrix (Figure 22). The most significant difference between the groups lay in the mean excerpt listening times of the secondary groupings (T2 and T3) in P2. Strikingly, the I group's T2 mean excerpt listening time was longer than their T1 listening time in both parts, although the difference was not significant. Compared to the less pronounced and considerably more linear differences in the F group's mean excerpt listening times over the three tests this was a very different set of performance data. It suggested that a multi stage analytic processing of the stimuli, in the manner that this task required, was not the approach that the I group would usually employ in musical sense-making.

The cognitive strategies of analytic vs holistic processing are the topic of both much research and much debate in experimental psychology. Dewey (1997) neatly summarizes the difference between the two as follows: "Analytic thinking involves understanding a system by thinking about its parts and how they work together to produce larger-scale effects. Holistic thinking involves understanding a system by sensing its large-scale patterns and reacting to them."

This definition is markedly similar to that encountered in section 2.4.3 in the description of the global precedence effect, and the local/global processing distinction. Indeed, research in both areas is founded on the same principles and stems from the original work of Navon (1977). However, in much of the associated literature, terms like analytic, holistic, global and local are used relatively interchangeably without clear definition or description of what they mean. This is a problem highlighted and discussed by Kimchi (1992). According to her review, authors use the term holistic processing in relation to at least two related but different processes:

> The first, which is considered to be more in the spirit of the Gestalt theory, refers to the primacy of wholistic [*sic*] properties in perception. In this usage, the terms *wholistic* and *global* are often used interchangeably to express the hypothesis that the initial information-processing step in the identification, discrimination, or classification of objects involves processing of wholistic properties rather than component properties… The other usage refers to the notion that the unitary whole, rather than its properties (whether wholistic or component), is the primary unit for processing. (p.25)

Accordingly, we can consider the terms local and global as somewhat interchangeable with the terms analytic and holistic. Reconsidering the task process in this study in light of these distinctions, it is apparent that the primary grouping can be completed holistically or analytically, using global or local components. The secondary grouping, however, seems to set up an automatic analytical context. If a participant's cognitive style is analytical from the outset, they use local components in each grouping. If it is holistic they might use a global component at T1, but are forced, by the requirement to make a secondary grouping, into using local components in T2 and T3.

Listeners with formal experience apparently had no difficulty in listening analytically to music - decomposing it into component properties - a skill commonly taught as part of formal musical training. It seems that listeners with informal experience, in line with behaviors observed in Ouimet, Foster & Hyde (2012), may have exhibited a greater tendency towards holistic processing in T1, but were forced to adopt a more analytic style in T2 and T3. This proposed shift in cognitive style offers a plausible explanation for the significant differences between groups in mean excerpt listening times, especially the observation that the I group's T2 times were longer than their T1 times. It might also account for the greater, although not significant, difference in perceptual error rates in the secondary groupings between groups. The results of Ouimet, Foster and Hyde (2012) suggest that global processing occurs significantly faster than local, that informal listeners were slower, albeit not significantly, in processing local components, and that formal listeners showed greater accuracy in local processing. Finally, the suggested differences in cognitive style – that the formal group tended to process more analytically from the outset – might offer an explanation as to the relatively minor role of melodic contour in the grouping decisions of the formal group, given that it is identified in the literature as a global feature (Mottron et al., 2000).

If the observed difference in mean excerpt listening times is indicative of different cognitive strategies at work in music listening, then apparently both can be applied with relatively equal levels of success. Questions naturally arise, though, as to what the hallmarks of these differing strategies might be, and which strategy is the more commonplace in everyday music listening.

6 CONCLUSION

The assumption is that life doesn't need to be navigated with lessons. You can just do it intuitively – Alain de Botton

6.1 Summary of Aims and Findings

This thesis set out to determine whether individuals employ a weighting of musical parameters in the processing of musical materials in a context analogous to everyday undirected music listening. It asked what style of parameter weighting might be occur: an approach based on bias or one based on parameter salience. It also aimed to investigate what differences were exhibited in the weighting of parameters and the processing of materials between groups comprised of individuals with differing levels of formal musical training.

Results demonstrated that individuals employed varied and idiosyncratic schema of parameter weight in the task, and that these were more likely due to individual bias than the salience of a parameter in a musical context. An extension of the employed weighting system also suggested that, in many cases, participants might be applying the same hierarchical schema of cue weight to the different melodic contexts presented in each part of the task.

Participant groups exhibited a number of key differences in task completion and processing: Groups employed collective schemas of parameter weight that featured different hierarchical and distributional organization. The formal group's most heavily weighted parameter was metrical and agogic accent, the informal group's was melodic contour. The groups' perceptual error rates were different both between and within-group for each part of the task. In the first part, where an even number of musical and auditory parameters were available for similarity judgements, the formal group had a perceptual advantage. In the second part, where auditory parameters outweighed musical ones, the informal group had an advantage. The formal group also exhibited an overall processing advantage, needing less time to complete each matrix than the informal group. The greatest difference between the groups was in the mean listening time per excerpt, where the informal group needed longer, particularly at the first of the secondary groupings.

6.2 Limitations of the Current Study

Despite the stated emphasis on providing participants with an everyday listening experience, it was necessary to present artificial tightly-controlled stimuli in this study. Due to this and the small sample size, the ability to generalize these findings to the wider population is somewhat limited. Furthermore, whilst much research has been undertaken on melodic similarity, it is questionable how analogous this kind of task is to real-time music listening. Whilst the method employed in this study was effective, further research, that would seek to generalize its findings, would need to be based on a larger sample and address the issue of more naturalistic stimulus materials, employing real or more realistic timbres, longer duration of stimuli and potentially moving into more common musical forms such as homophony, rather than monophony.

In the analysis of the results of this study, the weighting scheme that was employed worked reasonably well, but it is important to note that it *can* only be applied to these results. We in no way seek to imply that melodic contour is 27% more useful as a source of cues to informal music listeners in general music listening, for instance. Other areas of research have, however, developed more robust models of cue weight (see Toscano & McMurray, 2011) and it may be possible to do the same for music, allowing a much greater generalization of the results. Further to this, the chi-square statistic that was employed to test the goodness-of-fit of weighting distributions, was not ideal. When participants had lost points due to perceptual error, it tended to exaggerate the differences in the smaller sample. Fortunately, this did not result in any type II errors in our results, but this problem was noted. Future research, then, needs to find, adapt or develop more robust methods of cue weighting.

Finally, a major difference was in evidence between our stimuli and those employed in Lotto and Holt (2006) - that of psychophysical matching. Despite efforts to ensure that the parameters in the stimuli used in this study would be perceived as going through equal amounts of manipulation (see section 3.3 for measures taken), the stimuli were not submitted to perceptual rating prior to use in the experiment. Consequently, we have no way of knowing how equally the changes were perceived (see section 5.2). It can be argued that matching was unnecessary in our experiment - if the stimuli were unequally manipulated this should have presented a situation in which one parameter emerged as more salient than others, which, logically, should have led to its selection by all participants if salience were the main classification criterion. If bias were employed rather than salience, which was suggested by our results, then an equal or unequal starting point would not affect the outcome. However, it would allow greater understanding of results, and make for an interesting topic in itself, if future research employed psychophysical matching on the changes in musical stimuli prior to the use of those stimuli in a classification task.

6.3 Possible Directions for Future Research

As well as the concerns raised by the identified limitations, those of larger sample size, more naturalistic and psychophysically matched stimuli, several other questions have arisen as a result of this study that we propose as directions for future research.

The results of this study suggest individual schemas of cue weighting that are based on bias. How stable are these schemas? The extension of the weighting scheme suggested that participants were applying the same hierarchical schema to two different musical contexts (P1 and P2), but is this the case? Further research should test this hypothesis using a valid approach. Furthermore, is the schema stable? Would it change from day-to-day, so that on a different date and time a participant would apply a different schema to musical material with the same cues available? If the schema changed, what factors would affect this? We suggest the collection of similarity ratings on different occasions, with stimuli that offer the same parameters in different musical contexts, in order to test the stability of the cue weighting schema. Furthermore, the collection of information about music prefences, mood and personality would afford greater insight into any changes that may occur.

The prominent role of distortion in the F group’s P1 schema raised questions about the valence associated with grouping decisions. What role does valence play in grouping? Does a negatively valenced parameter outweigh a positively valenced parameter in a cue weighting schema? How much would this affect groupings in areas of possible advantage and disadvantage such as were exhibited in the responses to this study. We propose that future research would incorporate some measure with which to capture participant valence during grouping, so as to explore these questions more fully.

The data captured by this study suggested that while informal participants were capable of the analytic processing of musical stimuli, it may not be the cognitive strategy that they commonly employ. We suggest that future research could consider the role of cue weighting in experimental settings promoting other cognitive styles, such as fixed versus diverted attention tasks, to investigate whether a change in cognitive strategy results in a change in parameter weighting.

6.4 Final Thoughts

This study took a unique approach to an ongoing problem - to answer the question of how we listen to music. It focused its attention purely on music listeners and their cognitive decisions in an undirected context, in a conscious attempt to move away from the paradigm of directing participant cognition towards a specific contextual outcome. Its findings have important implications for music cognition research, in considering the role of individual bias in musical decision-making, in highlighting potential differences in cognitive strategies, and in informing the possible direction of future research. As such, it represents a constructive first step towards a more complete answer to the fundamental question with which it began.

7 REFERENCES

Addessi, A. R., & Caterina, R. (2000). Perceptual musical analysis: Segmentation and perception of tension. *Musicae Scientiae, 4*(1), 31-54.

Amer, T., Kalender, B., Hasher, L., Trehub, S. E., & Wong, Y. (2013). Do older professional musicians have cognitive advantages? *PLoS ONE 8*(8), e71630.

Balkwill, L. L., & Thompson, W. F. (1999). A cross-cultural investigation of the perception of emotion in music: Psychophysical and cultural cues. *Music Perception: An Interdisciplinary Journal 17*(1), 43-64.

Batt-Rawden, K., & DeNora, T. (2005). Music and informal learning in everyday life. *Music Education Research*, *7*(3), 289-304.

Benward, B., & Saker, M. (2008). *Music in Theory and Practice* (8th Ed. Vol. 1). McGraw-Hill.

Besson, M., Chobert, J., & Marie, C. (2011). Transfer of training between music and speech: common processing, attention, and memory. *Frontiers in psychology*, *2*(94).

Bigand, E., & Poulin-Charronnat, B. (2006). Are we "experienced listeners"? A review of the musical capacities that do not depend on formal musical training. *Cognition*, *100*(1), 100-130.

Bregman, A. S. (1994). *Auditory scene analysis: The perceptual organization of sound*. MIT press.

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2010). The perception of structural boundaries in polyphonic representations of Western popular music. *Musicae Scientiae, 14*(1 suppl), 115-155.

Carey, D., Rosen, S., Krishnan, S., Pearce, M. T., Shepherd, A., Aydelott, J., & Dick, F. (2015). Generality and specificity in the effects of musical expertise on perception and cognition. *Cognition*, *137*, 81-105.

Cook, N. (1994). *A Guide to Musical Analysis.* Oxford University Press.

Copland, A. (2011). *What to Listen for in Music*. Penguin.

Deliège, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's grouping preference rules. *Music Perception: An Interdisciplinary Journal, 4*(4), 325-360.

Deliège, I. (1989). A perceptual approach to contemporary musical forms. *Contemporary Music Review*, *4*(1), 213-230.

Deliège, I. (1996). Cue abstraction as a component of categorisation processes in music listening. *Psychology of Music*, *24*(2), 131-156.

Deliège, I. (2001). Similarity Perception↔ Categorization↔ Cue Abstraction. *Music Perception: An Interdisciplinary Journal*, *18*(3), 233-243.

Dewey, R. A. (2007). *Psychology: an introduction*. Russ Dewey.

Drabkin, W. (2015). “Motif” In *Grove Music Online*. Oxford University Press. Retrieved from http://www.oxfordmusiconline.com/subscriber/article/grove/music/19221

Eerola, T., Järvinen, T., Louhivuori, J., & Toiviainen, P. (2001). Statistical features and perceived similarity of folk melodies. *Music Perception: An Interdisciplinary Journal*, *18*(3), 275-296.

Eitan, Z., & Granot, R. Y. (2009). Primary versus secondary musical parameters and the classification of melodic motives. *Musicae Scientiae,13*(1 suppl), 139-179.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature, 415*(6870), 429-433.

Florentine, M., Popper, A. N., & Fay, R. R. (2011). *Loudness*. Springer.

Folkestad, G. (2006). Formal and informal learning situations or practices vs formal and informal ways of learning. *British journal of music education*, *23*(2), 135-145.

Francis, A. L., Baldwin, K., & Nusbaum, H. C. (2000). Effects of training on attention to acoustic cues. *Perception & Psychophysics*, *62*(8), 1668-1680.

Fujioka, T., Trainor, L., Ross, B., Kakigi, R., & Pantev, C. (2004). Musical training enhances automatic encoding of melodic contour and interval structure. *Journal of Cognitive Neuroscience*, *16*(6), 1010-1021.

Garson, G. D. (2013). Fundamentals of hierarchical linear and multilevel modeling. *Hierarchical linear modeling: guide and applications.* Sage Publications Inc.

Granot, R. Y., & Eitan, Z. (2011). Musical tension and the interaction of dynamic auditory parameters. *Music Perception: An Interdisciplinary Journal*, *28*(3), 219-246.

Greasley, A. E., & Lamont, A. (2011). Exploring engagement with music in everyday life using experience sampling methodology. *Musicae Scientiae*,*15*(1), 45-71.

Hajda, J. M., Kendall, R. A., Carterette, E. C., & Harshberger, M. L. (1997). Methodological issues in timbre research. *Perception and Cognition of Music* (253-306). Psychology Press.

Handel, S. (1993). *Listening: An introduction to the perception of auditory events*. The MIT Press.

Holt, L. L., & Lotto, A. J. (2006). Cue weighting in auditory categorization: Implications for first and second language acquisition. *The Journal of the Acoustical Society of America*, *119*(5), 3059-3071.

Hillenbrand, J. M., Clark, M. J., & Houde, R. A. (2000). Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America*, *108*(6), 3013-3022.

Horn, D., Laing, D., Oliver, P., & Wicke, P. (2003). *Continuum Encyclopedia of Popular Music of the World Part 1 Performance and Production* (Vol. 2). Bloomsbury Academic.

Huron, D. (2002). Listening Styles and Listening Strategies. Retrieved from http://www.musiccog.ohio-state.edu/Huron/Talks/SMT.2002/handout.html

Jacobs, R. A. (1999). Optimal integration of texture and motion cues to depth. *Vision Research*, *39*(21), 3621-3629.

Juslin, P. N. (2000). Cue utilization in communication of emotion in music performance: relating performance to perception. *Journal of Experimental Psychology: Human Perception and Performance, 26*(6), 1797.

Juslin, P. N., Liljeström, S., Västfjäll, D., Barradas, G., & Silva, A. (2008). An experience sampling study of emotional reactions to music: listener, music, and situation. *Emotion*, *8*(5), 668.

Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: a critical review. *Psychological bulletin*, *112*(1), 24-38.

Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *The Journal of the Acoustical Society of America*, *118*(2), 1038-1054.

Koelsch, S., Schmidt, B. H., & Kansok, J. (2002). Effects of musical expertise on the early right anterior negativity: An event-related brain potential study. *Psychophysiology*, *39*(5), 657-663.

König, O., Schaette, R., Kempter, R., & Gross, M. (2006). Course of hearing loss and occurrence of tinnitus. *Hearing research, 221*(1), 59-64.

Kuhl, P. K. (2004). Early language acquisition: cracking the speech code. *Nature Reviews Neuroscience*, *5*(11), 831-843.

Lamont, A., & Dibben, N. (2001). Motivic structure and the perception of similarity. *Music Perception: An Interdisciplinary Journal*, *18*(3), 245-274.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *Journal of the Optical Society of America A*, *18*(9), 2307-2320.

Lerdahl, F., & Jackendoff, R. (1985). *A generative theory of tonal music*. MIT press.

Levitin, D. J., & Menon, V. (2003). Musical structure is processed in "language" areas of the brain: a possible role for Brodmann Area 47 in temporal coherence. *Neuroimage, 20*(4), 2142-2152.

Levitin, D. J., & Tirovolas, A. K. (2009). Current advances in the cognitive neuroscience of music. *Annals of the New York Academy of Sciences*, *1156*(1), 211-231.

Luo, X., Masterson, M. E., & Wu, C. C. (2014). Contour identification with pitch and loudness cues using cochlear implants. *The Journal of the Acoustical Society of America*, *135*(1), EL8-14.

McAdams, S. (1999). Perspectives on the contribution of timbre to musical structure. *Computer Music Journal*, *23*(3), 85-102.

McAdams, S., Vieillard, S., Houix, O., & Reynolds, R. (2004). Perception of musical similarity among contemporary thematic materials in two instrumentations. *Music Perception: An Interdisciplinary Journal*, *22*(2), 207-237.

Mikutta, C. A., Maissen, G., Altorfer, A., Strik, W., & Koenig, T. (2014). Professional musicians listen differently to music. *Neuroscience*, *268*, 102-111.

Mottron, L., Peretz, I., & Menard, E. (2000). Local and Global Processing of Music in High-functioning Persons with Autism: Beyond Central Coherence? *Journal of Child Psychology and Psychiatry*, *41*(8), 1057-1065.

Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive Psychology*, *9*(3), 353-383.

North, A. C., Hargreaves, D. J., & Hargreaves, J. J. (2004). Uses of music in everyday life. *Music Perception: An Interdisciplinary Journal*, *22*(1), 41-77.

Ouimet, T., Foster, N. E., & Hyde, K. L. (2012). Auditory global-local processing: Effects of attention and musical experience. *The Journal of the Acoustical Society of America*, *132*(4), 2536-2544.

Owsinski, B. (2013). *The mixing engineer's handbook*. Cengage Learning.

Parsons, D. (1975). *The directory of tunes and musical themes.* Spencer Brown.

Patel, A. D. (2007). *Music, Language, and the Brain*. Oxford University Press.

Pitt, M. A. (1994). Perception of pitch and timbre by musically trained and untrained listeners. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 976.

Poulin-Charronnat, B., Bigand, E., Lalitte, P., Madurell, F., Vieillard, S., & McAdams, S. (2004). Effects of a change in instrumentation on the recognition of musical materials. *Music Perception: An Interdisciplinary Journal*, *22*(2), 239-263.

Reybrouck, M. (2009). Similarity perception as a cognitive tool for musical sense-making: deictic and ecological claims. *Musicae Scientiae*, *13*(1 suppl), 99-118.

Reybrouck, M. (2010). Music cognition and real-time listening: denotation, cue abstraction, route description and cognitive maps. *Musicae scientiae*,*14*(2 suppl), 187-202.

Robinson, K., & Patterson, R. D. (1995). The duration required to identify the instrument, the octave, or the pitch chroma of a musical note. *Music Perception: An Interdisciplinary Journal*, *13*(1), 1-15.

Rumsey, F., & McCormick, T. (2006). *Sound and Recording: An Introduction*. Taylor & Francis.

Schellenberg, E. G., & Moreno, S. (2010). Music lessons, pitch processing, and g. *Psychology of Music*, *38*(2), 209-221.

Shepard, R. N. (1999). Cognitive Psychology and Music. *Music, Cognition and Computerized Sound* (21 – 36). The MIT Press.

Sloboda, J. A., O'Neill, S. A., & Ivaldi, A. (2001). Functions of music in everyday life: An exploratory study using the experience sampling method. *Musicae scientiae*, *5*(1), 9-32.

Stoesz, B. M., Jakobson, L. S., Kilgour, A. R., & Lewycky, S. T. (2007). Local processing advantage in musicians: Evidence from disembedding and constructional tasks. *Music Perception: An Interdisciplinary Journal*, *25*(2), 153-165.

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434-464.

Trainor, L. J., Desjardins, R. N., & Rockel, C. (1999). A Comparison of Contour and Interval Processing in Musicians and Nonmusicians Using Event-Related Potentials. *Australian Journal of Psychology*, *51*(3), 147-153.

White, J. D. (1984). *The analysis of music*. Scarecrow Press.

Ziv, N., & Eitan, Z. (2007). Themes as prototypes: Similarity judgments and categorization tasks in musical contexts. *Musicae Scientiae*, *11*(1 suppl), 99-133.

Index of Appendices:

APPENDIX A IMAGES OF THE EXPERIMENTAL INTERFACE

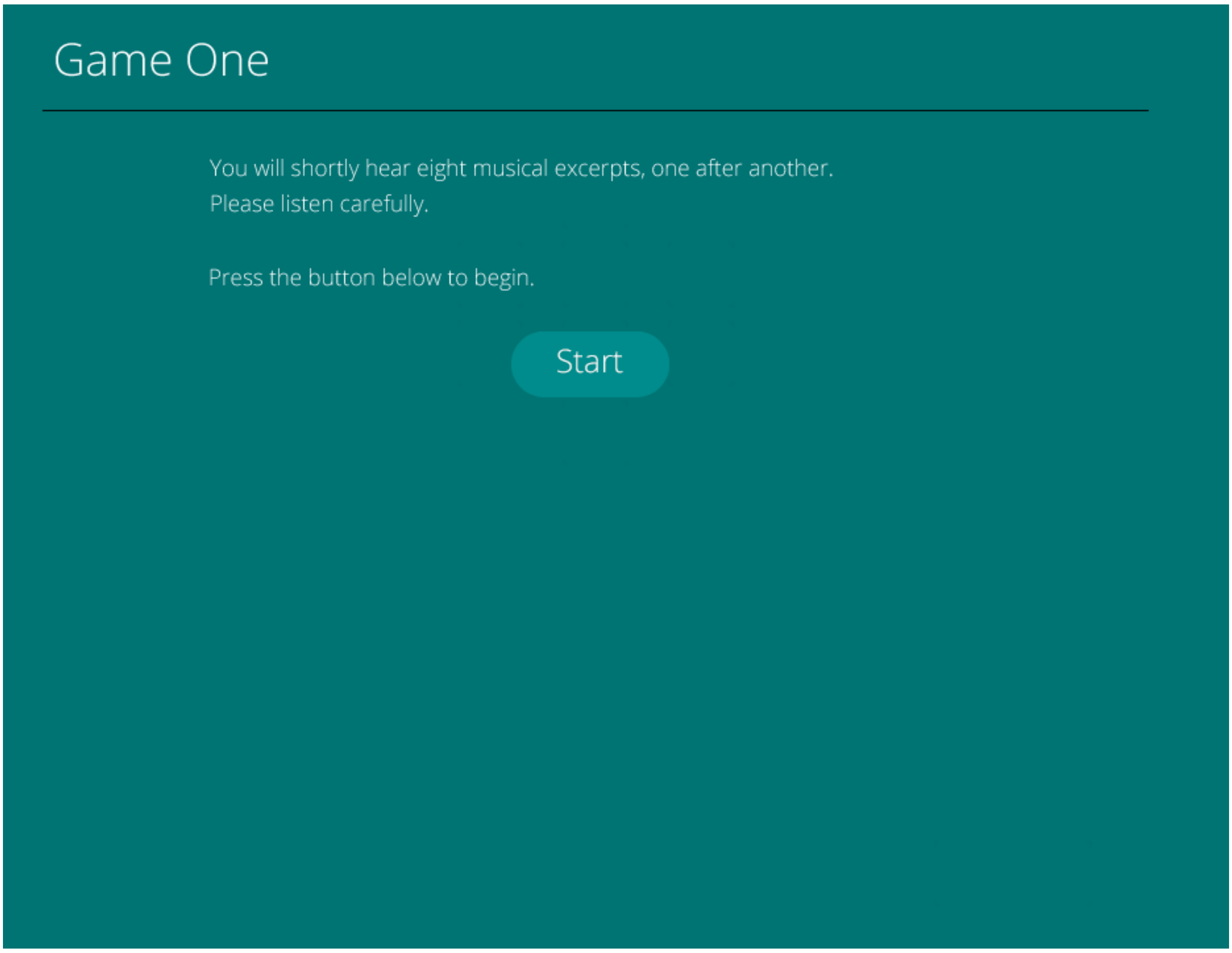


Fig 1. The starting screen of the main experimental interface. Participants press Start to begin.

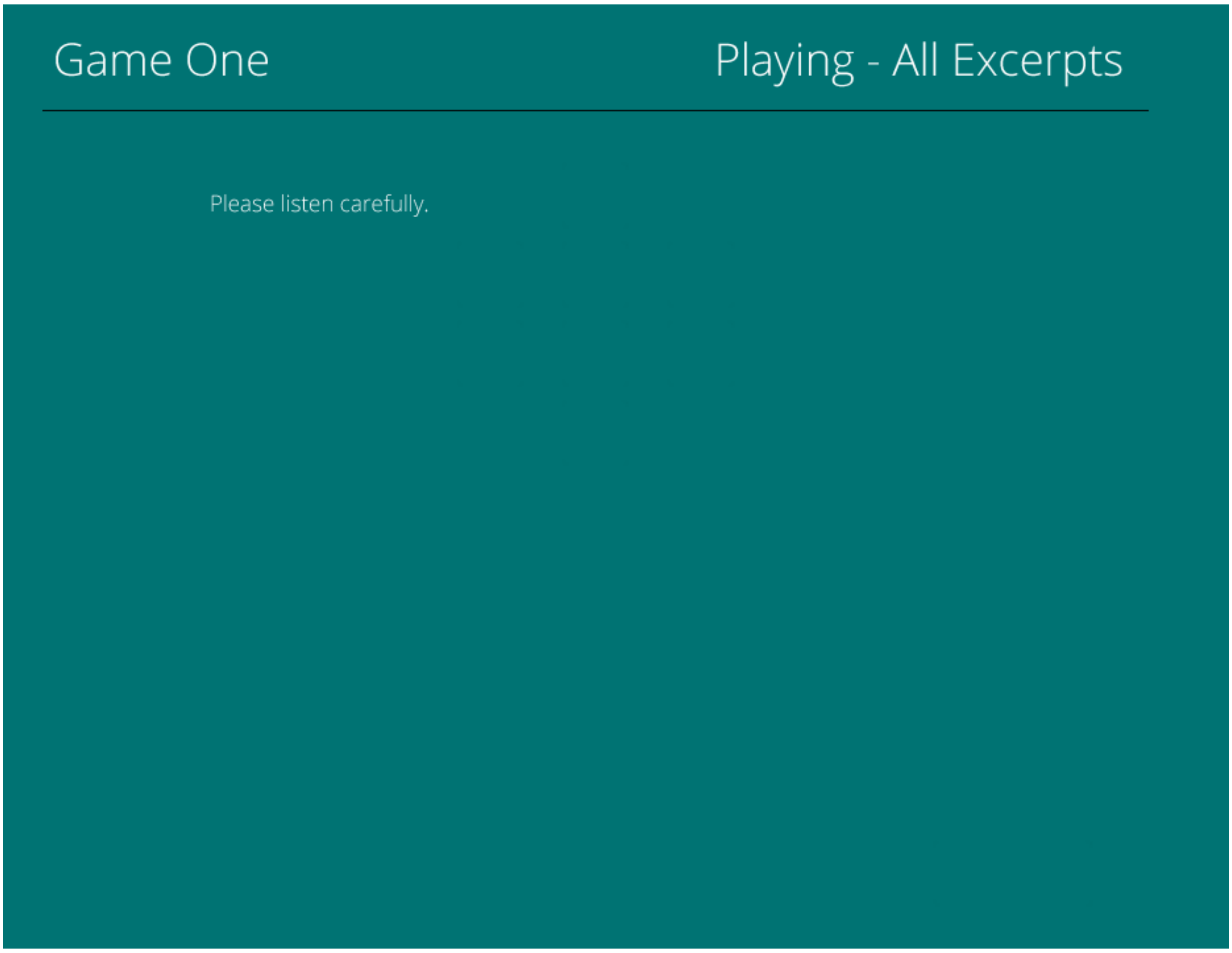


Fig 2. The screen was blank during priming.

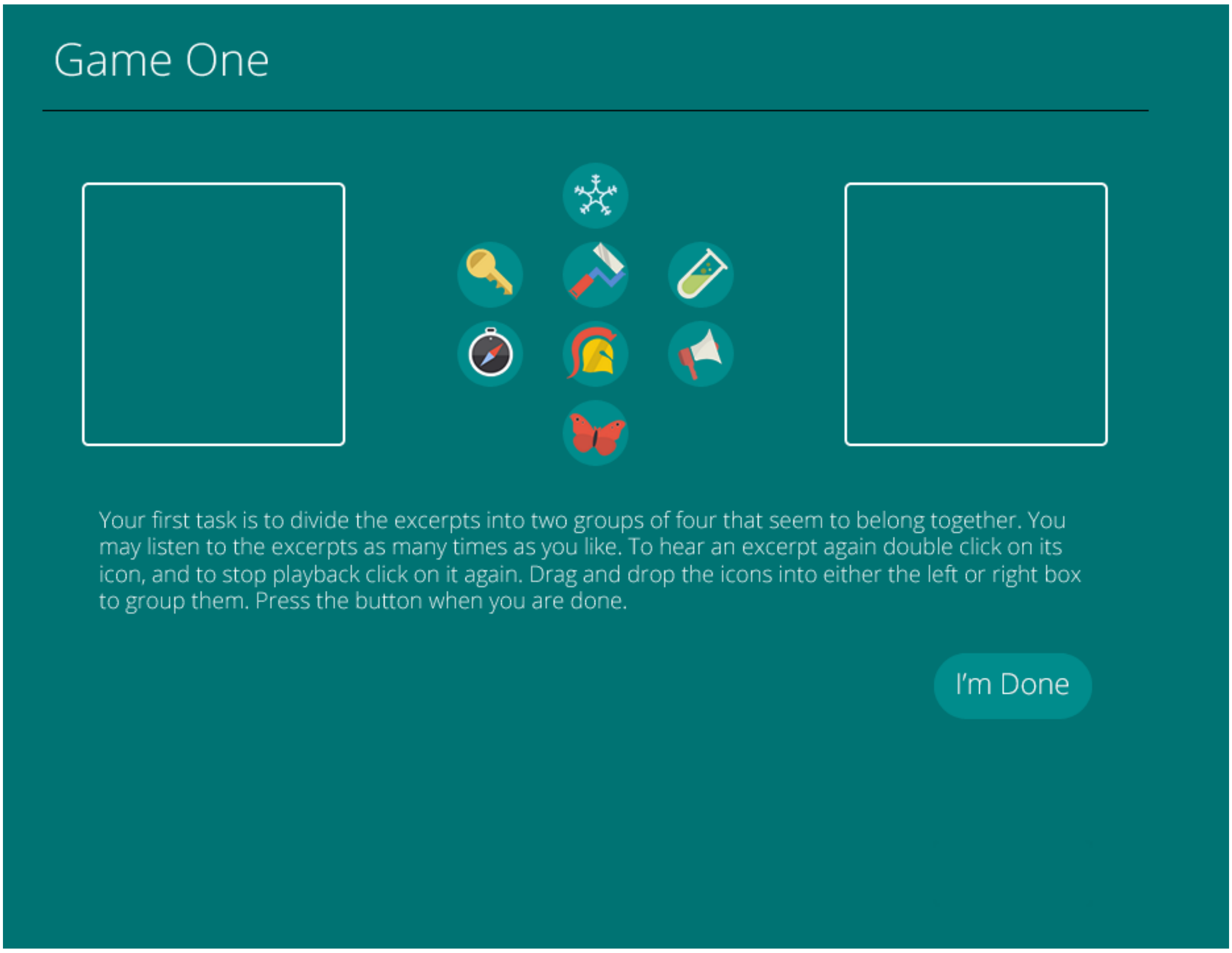


Fig 3. Post-priming participants were presented with the primary grouping task.

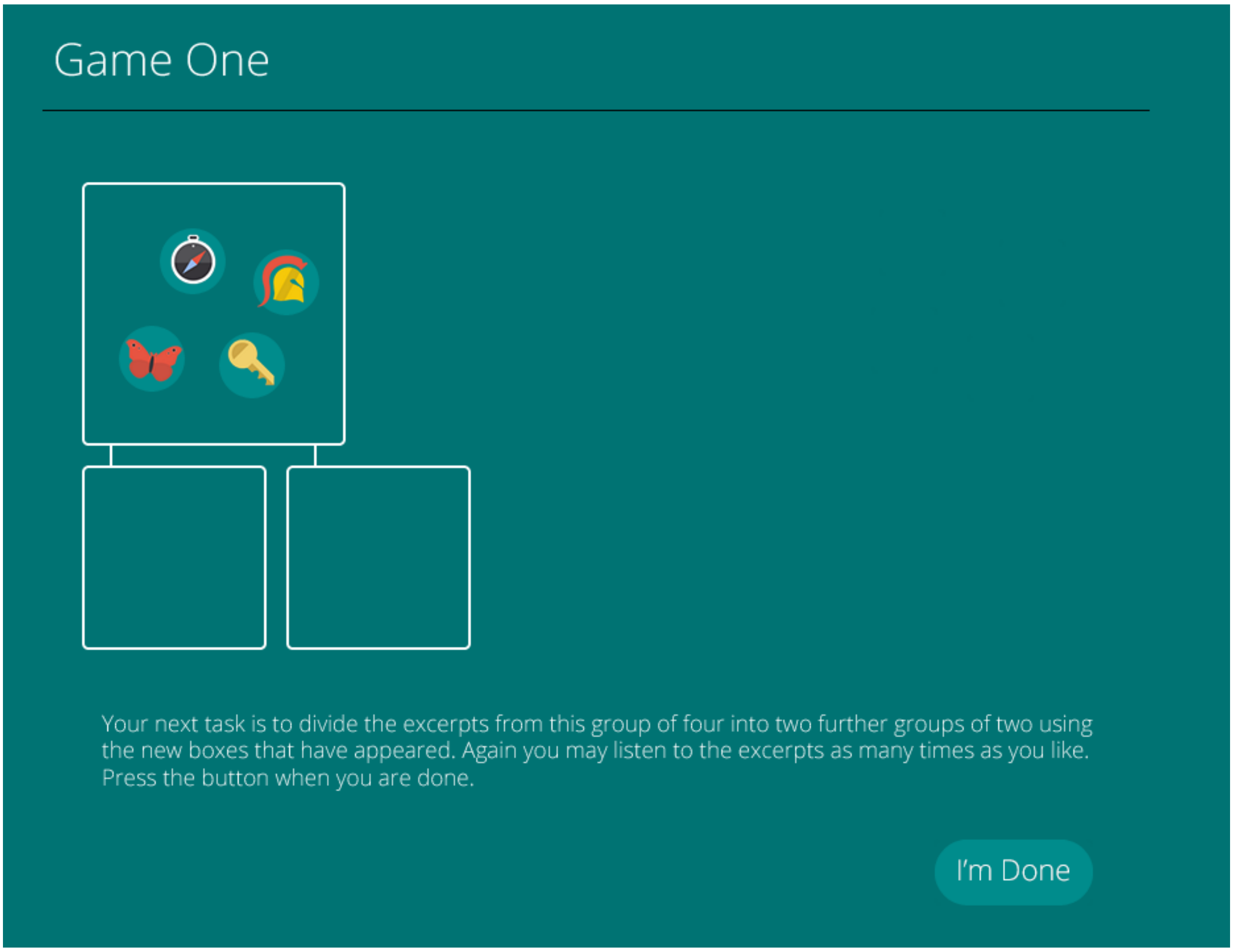
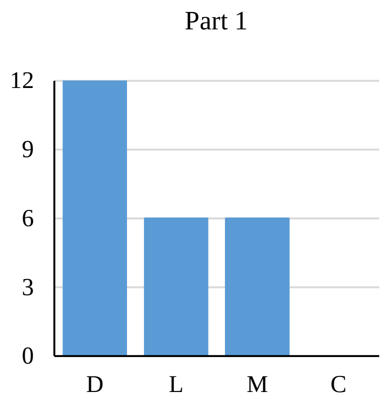


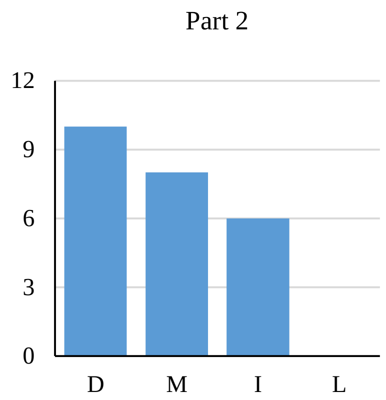
Fig 4. Once the primary grouping was successfully completed, participants progress to the first secondary grouping.

APPENDIX B INDIVIDUAL SCHEMAS

Participant 1 - Formal Group

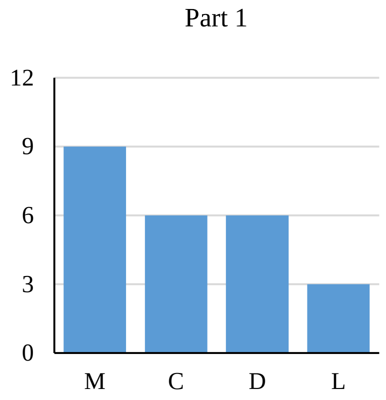


$\chi^2(3) = 12$, $p < .01$

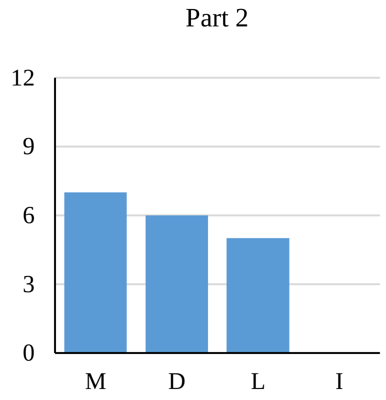


$\chi^2(3) = 9.33$, $p < .05$

Participant 2 - Formal Group

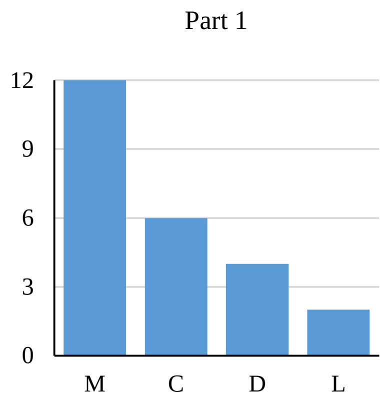


$\chi^2(3) = 3$, $p > .05$

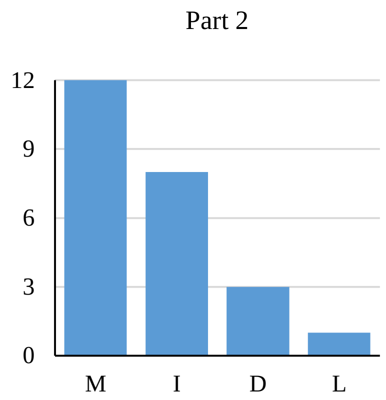


$\chi^2(3) = 6.44$, $p > .05$

Participant 3 - Formal Group

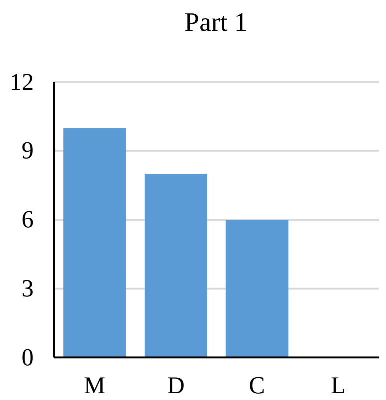


$\chi^2(3) = 9.33$, $p < .05$

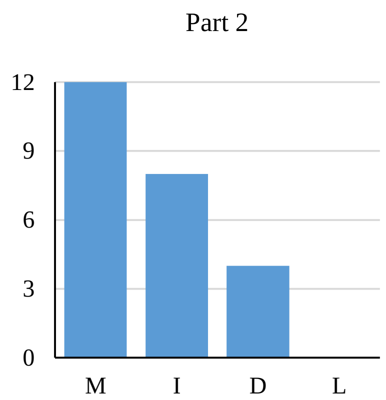


$\chi^2(3) = 12.33$, $p < .01$

Participant 4 - Formal Group

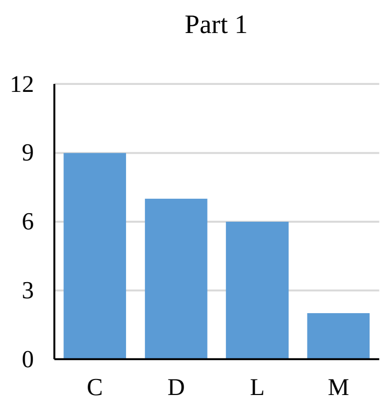


$\chi^2(3) = 9.33$, $p < .05$

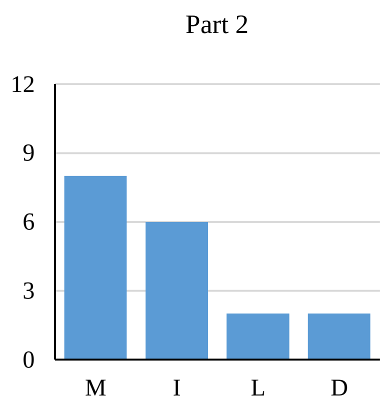


$\chi^2(3) = 13.33$, $p < .01$

Participant 5 - Formal Group

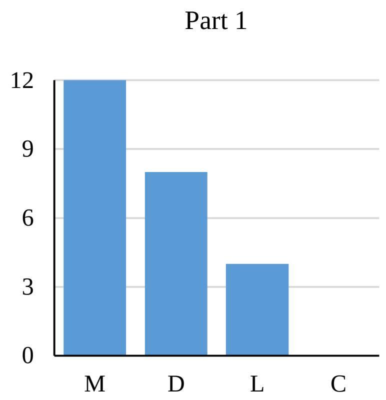


$\chi^2(3) = 4.33$, $p > .05$

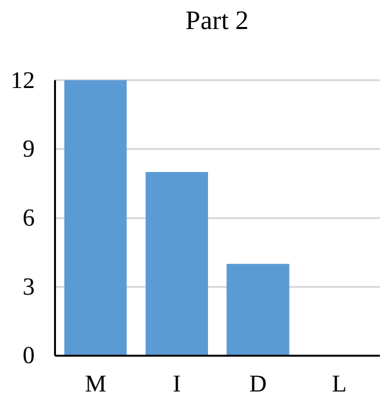


$\chi^2(3) = 6$, $p > .05$

Participant 6 - Formal Group

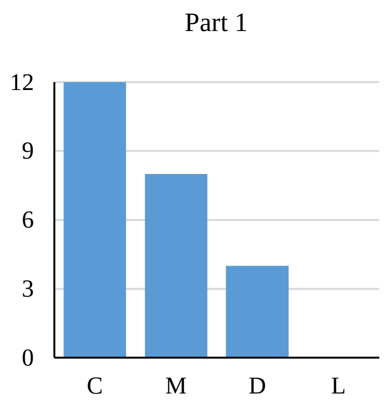


$\chi^2(3) = 13.33$, $p < .01$

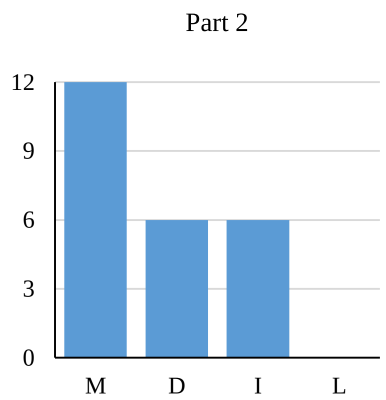


$\chi^2(3) = 13.33$, $p < .01$

Participant 7 - Formal Group

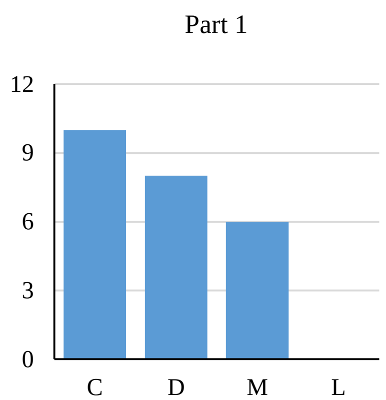


$\chi^2(3) = 13.33$, $p < .01$

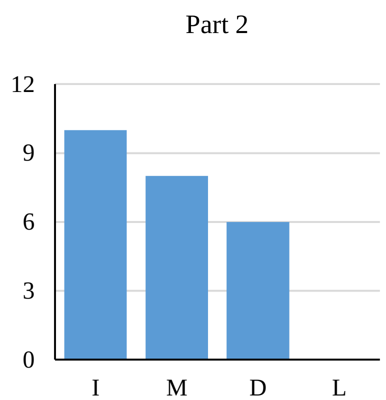


$\chi^2(3) = 12$, $p < .01$

Participant 8 - Formal Group

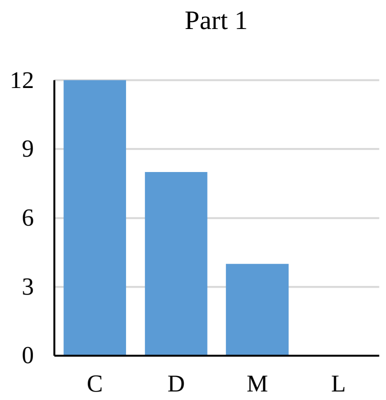


$\chi^2(3) = 9.33$, $p < .05$

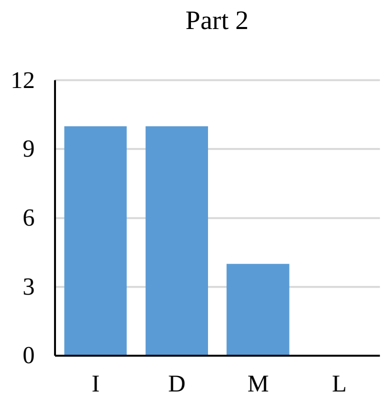


$\chi^2(3) = 9.33$, $p < .05$

Participant 9 - Formal Group

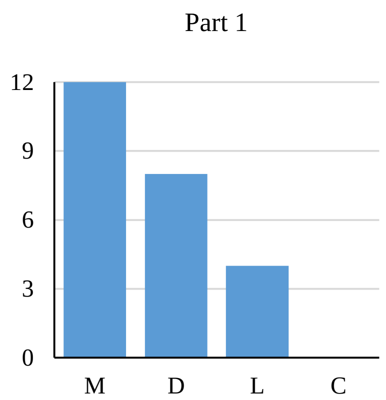


$\chi^2(3) = 13.33$, $p < .01$

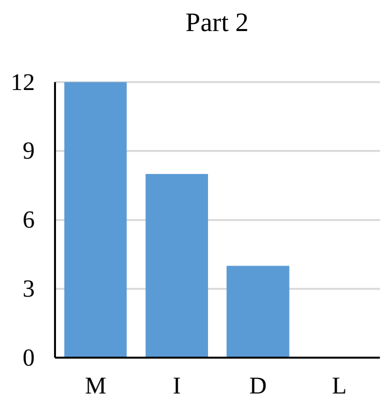


$\chi^2(3) = 12$, $p < .01$

Participant 10 - Formal Group

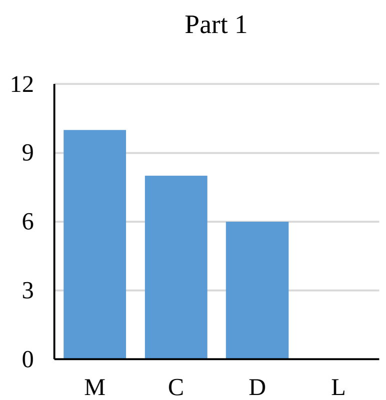


$\chi^2(3) = 13.33$, $p < .01$

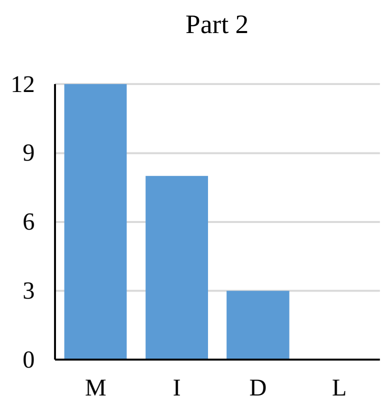


$\chi^2(3) = 13.33$, $p < .01$

Participant 11 - Formal Group

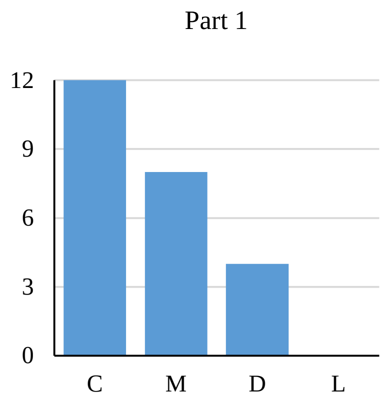


$\chi^2(3) = 9.33$, $p < .05$

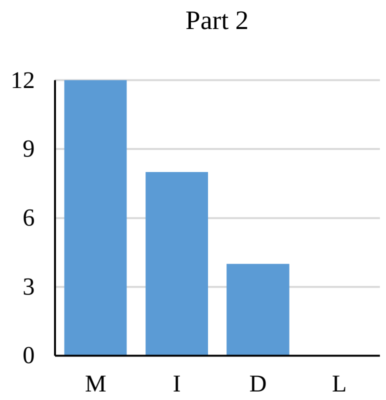


$\chi^2(3) = 14.74$, $p < .01$

Participant 12 - Formal Group

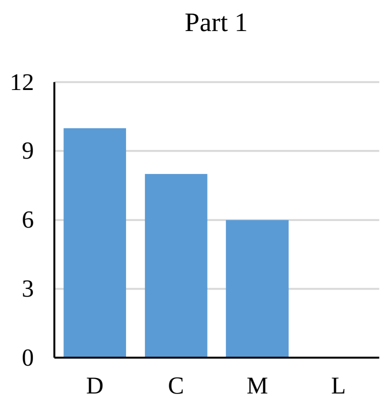


$\chi^2(3) = 13.33$, $p < .01$

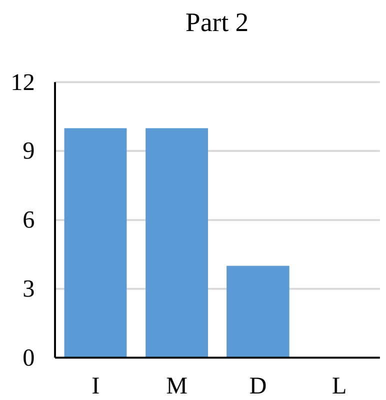


$\chi^2(3) = 13.33$, $p < .01$

Participant 13 - Formal Group

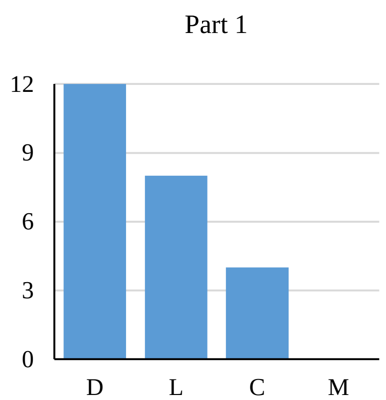


$\chi^2(3) = 9.33$, $p < .05$

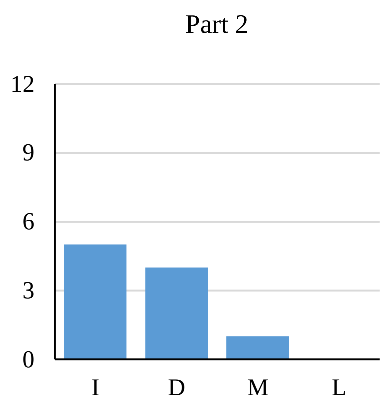


$\chi^2(3) = 12$, $p < .01$

Participant 14 - Formal Group

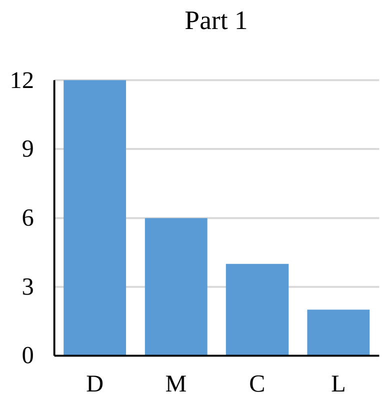


$\chi^2(3) = 13.33$, $p < .01$

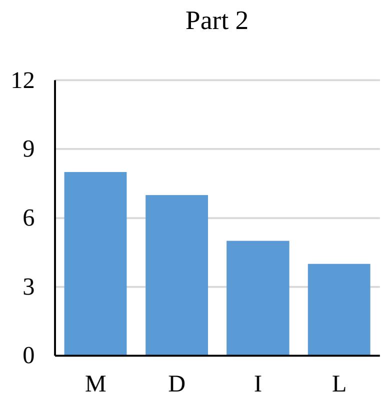


$\chi^2(3) = 6.8$, $p > .05$

Participant 15 - Formal Group

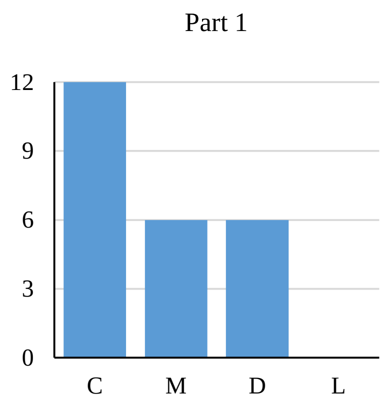


$\chi^2(3) = 9.33$, $p < .05$

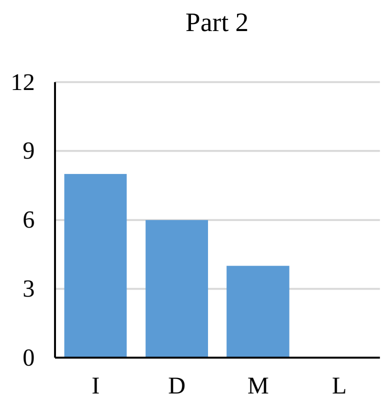


$\chi^2(3) = 1.67$, $p > .05$

Participant 1 - Informal Group

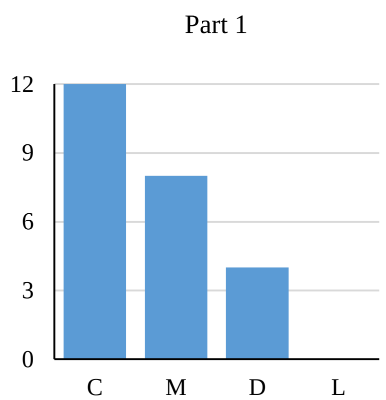


$\chi^2(3) = 12$, $p < .01$

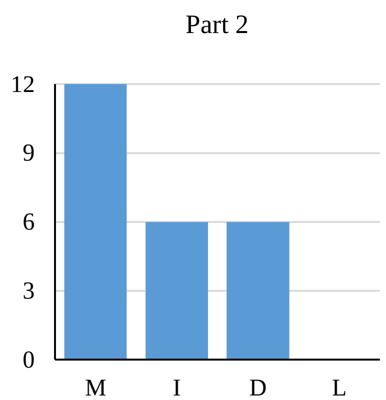


$\chi^2(3) = 7.78$, $p > .05$

Participant 2 - Informal Group

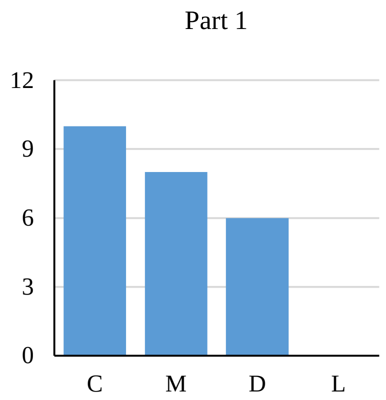


$\chi^2(3) = 13.33$, $p < .01$

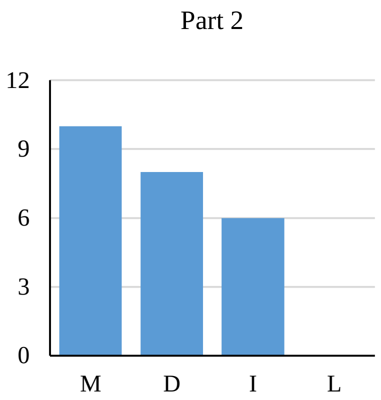


$\chi^2(3) = 12$, $p < .01$

Participant 3 - Informal Group

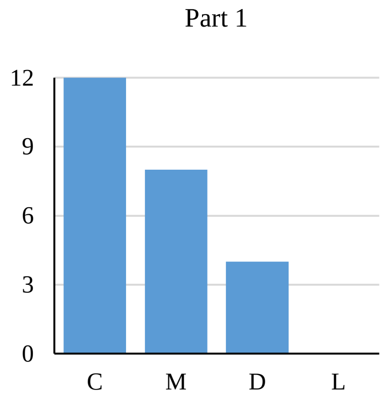


$\chi^2(3) = 9.33$, $p < .05$

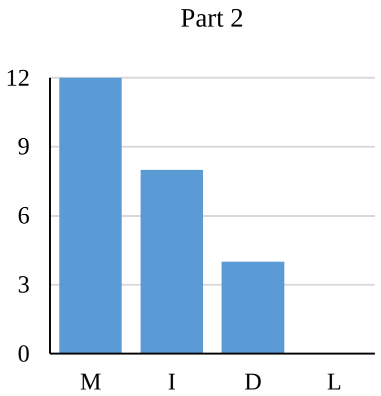


$\chi^2(3) = 9.33$, $p < .05$

Participant 4 - Informal Group

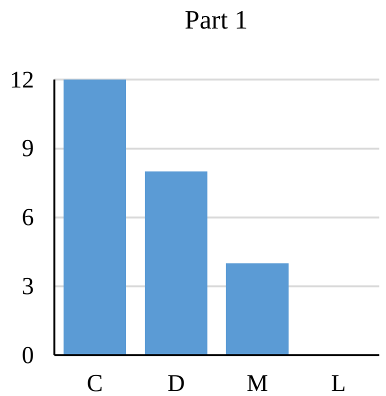


$\chi^2(3) = 13.33$, $p < .01$

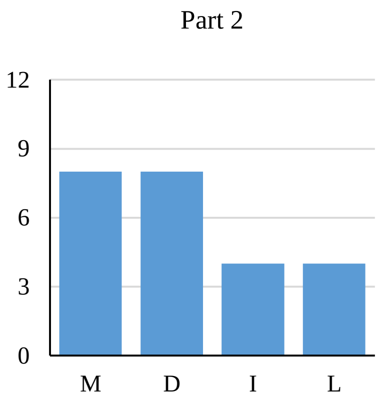


$\chi^2(3) = 13.33$, $p < .01$

Participant 5 - Informal Group



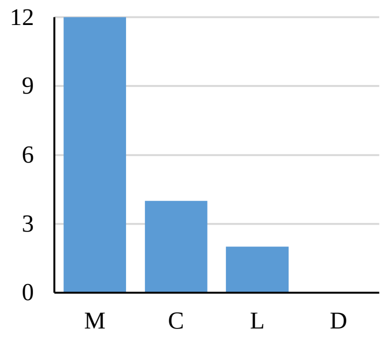
$\chi^2(3) = 13.33$, $p < .01$



$\chi^2(3) = 2.67$, $p > .05$

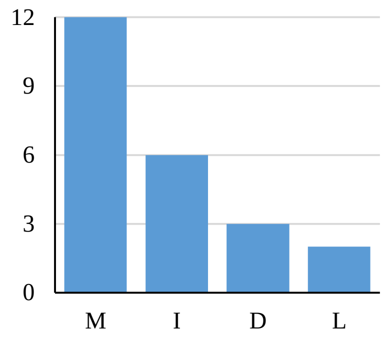
Participant 6 - Informal Group

Part 1



$\chi^2(3) = 18.44$, $p < .001$

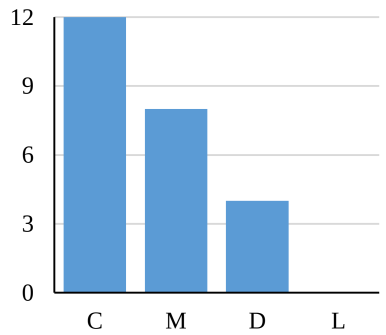
Part 2



$\chi^2(3) = 10.56$, $p < .05$

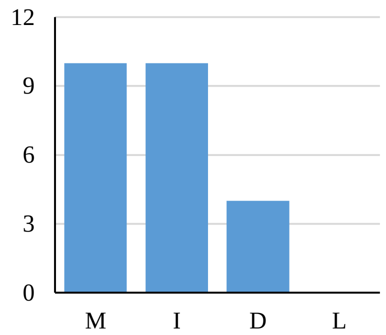
Participant 7 - Informal Group

Part 1



$\chi^2(3) = 13.33$, $p < .01$

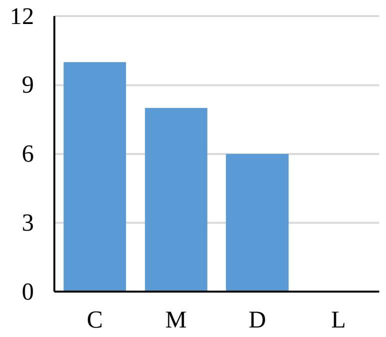
Part 2



$\chi^2(3) = 12$, $p < .01$

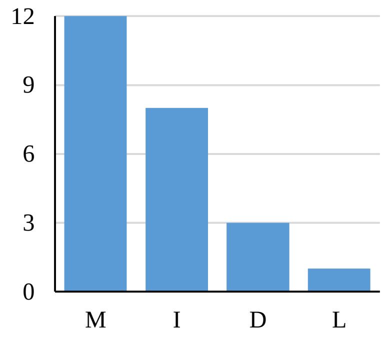
Participant 8 - Informal Group

Part 1



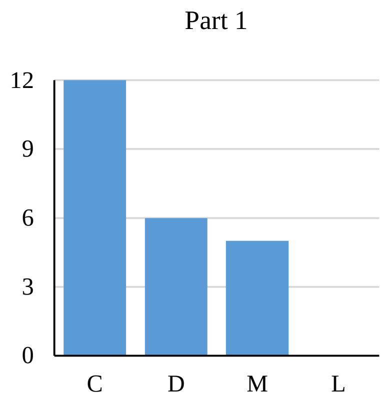
$\chi^2(3) = 9.33$, $p < .05$

Part 2

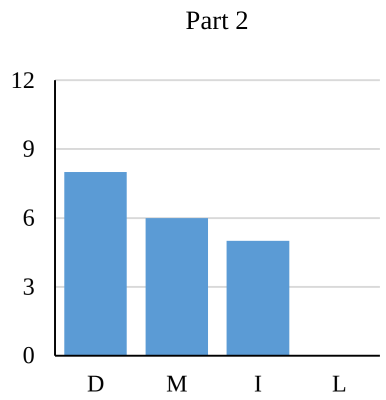


$\chi^2(3) = 12.33$, $p < .01$

Participant 9 - Informal Group

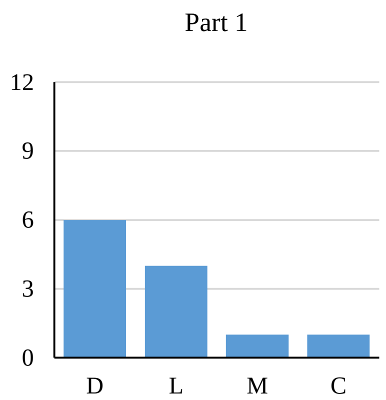


$\chi^2(3) = 12.65$, $p < .01$

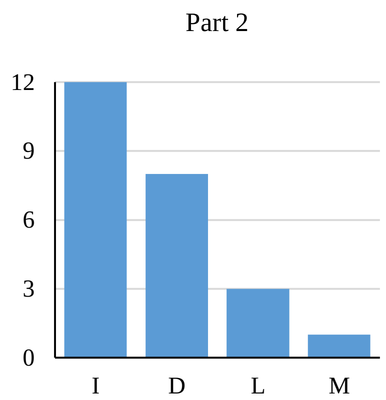


$\chi^2(3) = 7.32$, $p > .05$

Participant 10 - Informal Group

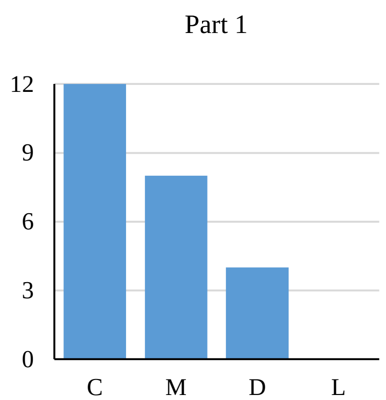


$\chi^2(3) = 6$, $p > .05$

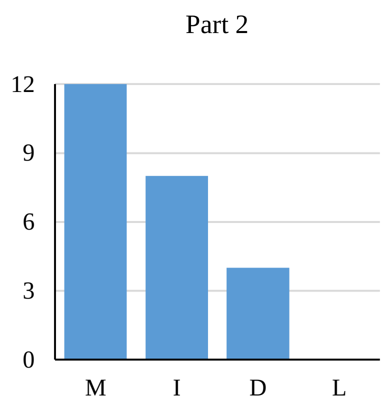


$\chi^2(3) = 12.33$, $p < .01$

Participant 11 - Informal Group

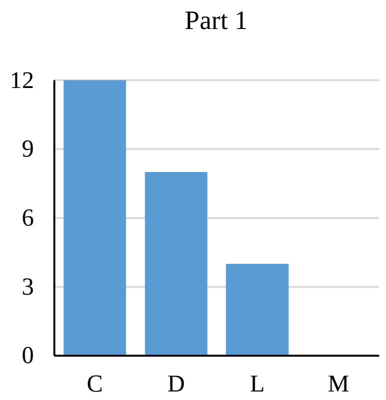


$\chi^2(3) = 13.33$, $p < .01$

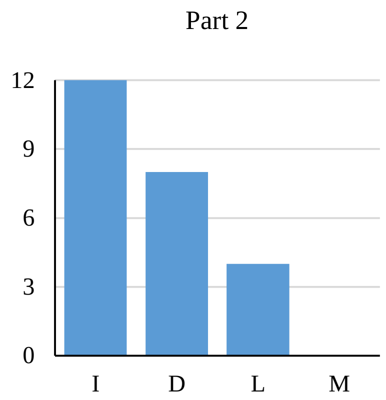


$\chi^2(3) = 13.33$, $p < .01$

Participant 12 - Informal Group

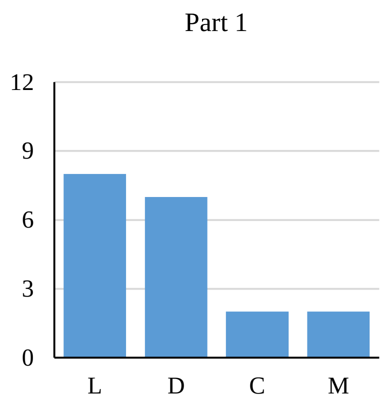


$\chi^2(3) = 13.33$, $p < .01$

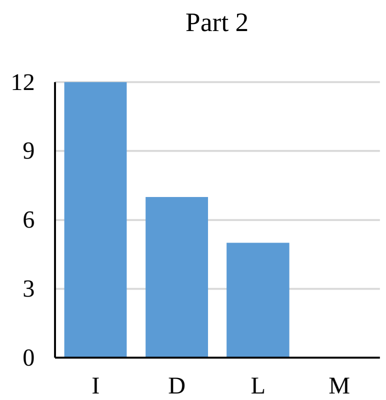


$\chi^2(3) = 13.33$, $p < .01$

Participant 13 - Informal Group

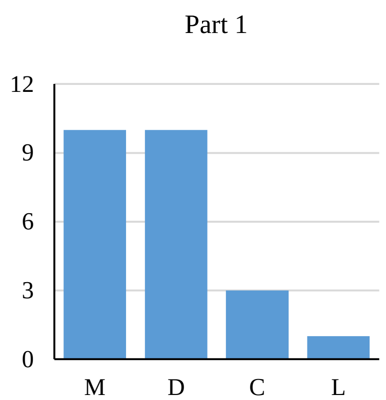


$\chi^2(3) = 6.474$, $p > .05$

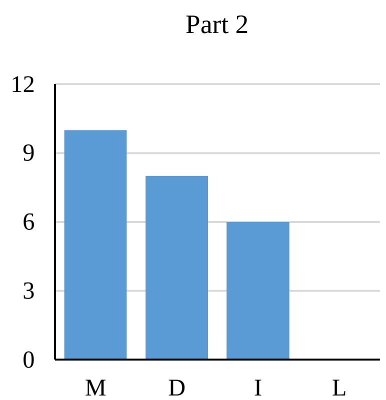


$\chi^2(3) = 12.33$, $p < .01$

Participant 14 - Informal Group



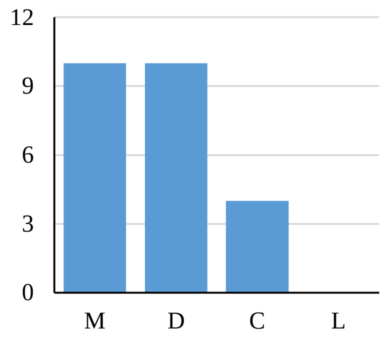
$\chi^2(3) = 11$, $p < .05$



$\chi^2(3) = 9.33$, $p < .05$

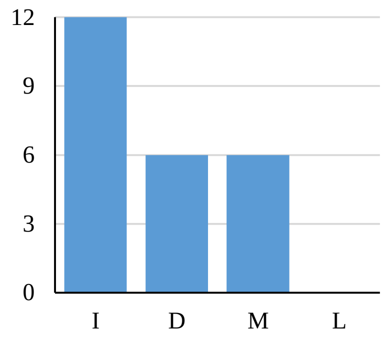
Participant 15 - Informal Group

Part 1



$\chi^2(3) = 9.33$, $p < .05$

Part 2



$\chi^2(3) = 12$, $p < .01$