

WESTERN SYDNEY UNIVERSITY

DOCTORAL THESIS

---

**The Cognition of Harmonic Tonality  
in Microtonal Scales**

---

*Author:*  
Lillian M. Hearne

*Supervisors:*  
Dr. Andrew J. Milne  
Prof. Roger T. Dean

*A thesis submitted in fulfillment of the requirements  
for the degree of Doctor of Philosophy*

*in the*

Music Cognition and Action Group  
The MARCS Institute

2020



## *Acknowledgements*

First, I acknowledge the traditional custodians of the land upon which I completed my studies: The Darug, Tharawal (also historically referred to as Dharawal), and Eora peoples, and pay my respects to Elders past, present, and future.

I would next like to acknowledge the guidance and help of my supervisors Andrew Milne and Roger Dean, who were a pleasure to work with. They were always available to talk when in the office and responded to emails and returned comments on drafts very promptly, and I never felt left in the dark.

I'd also like to thank my colleagues from the lab for their help and advice and for their participation in my experiments. My colleague and housemate Ian Colley deserves a special mention here, for participating in every one of my experiments and providing valuable advice with using *Max* and *R* as well as in completing a PhD in general. My other housemates Rémi Marchand and Juan A. M. Fuentes and my ex-housemate Patrick Blown also deserve acknowledgement for their help with MATLAB and *LaTeX*, and my colleague Patti Nijhuis and my mum, Sonja Hearne for listening to my venting about computer issues and such. I'd also like to thank the technical team at MARCS for their help and patience with these issues, and the admin staff for their support throughout.

I'd like to acknowledge the members of my musical ensembles Helix Quartet, Cone of Confusion, The Lancer Band, Sydney Harmony and St Mary's Cathedral Choir who participated in multiple of my experiments; everyone at MARCS, or in AMPS, SMPC, ICMPC and MCM, who watched my presentations and asked questions or discussed my work afterwards; to Marcus Pearce and Psyche Loui and the students in their labs for having me discuss my work with them; and to Paul Erlich, Margo Schulter, Jacob Barton, Juhani Nuorvala, Mike Battaglia, Kite Giedraitis and other mentors and contemporaries in The Xenharmonic Alliance.

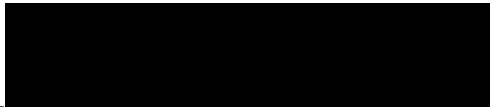
I'd like to thank my friends John Nicholls, for proofreading one of my papers, and Rachel Singer, for proofreading my papers and thesis; my girlfriend Indigo Terayama for providing helpful distraction in the final week before submission (and continued support and encouragement in the months following); and everyone I've spoken to who asked about my research and let me get into any sort of detail!

Finally, I'd like to thank Carol Krumhansl and Emery Schubert for their helpful reviews of this thesis.

## Statement of Authenticity

The work presented in this thesis is, to the best of my knowledge and belief, original except as acknowledged in the text. I hereby declare that I have not submitted this material, either in full or in part, for a degree at this or any other institution

Signed:

A solid black rectangular box redacting the signature of the author.

---

# Contents

<b>Acknowledgements</b>	<b>i</b>
<b>Statement of Authenticity</b>	<b>ii</b>
<b>Abstract</b>	<b>xvi</b>
<b>1 Background</b>	<b>1</b>
1.1 Introduction . . . . .	1
1.1.1 Tonality and Scale . . . . .	1
1.1.2 Thesis Summary . . . . .	3
1.2 Background . . . . .	4
1.2.1 Sound, spectrum, pitch and tuning . . . . .	4
1.2.2 Consonance and Affinity . . . . .	7
1.2.3 Probe Tone Experiments . . . . .	10
The context: presence or absence of tonal cues . . . . .	13
The dependent variable: participant ratings . . . . .	15
Participants: musicians or non-musicians . . . . .	16
1.2.4 Probe Chord Experiments . . . . .	16
Chord type and inversion . . . . .	20
Tension and stability . . . . .	21
1.2.5 The cognition of music in novel/microtonal tuning systems . . . . .	23
<b>2 Experiments 1 &amp; 2 – Tones</b>	<b>26</b>
2.1 Introduction . . . . .	26
2.2 Overview of the Experiments and Models . . . . .	27
2.2.1 Experimental design . . . . .	27
2.2.2 Hypotheses . . . . .	29
Experiment 1 . . . . .	29
Experiment 2 . . . . .	30

2.2.3	Analysis . . . . .	31
	Bayesian ordinal mixed effects models . . . . .	31
	Test for tonal hierarchies . . . . .	33
	Comparison of fit and stability ratings (H1) . . . . .	34
	Descriptive model . . . . .	34
	Testing H2 & 3 with Bayesian ordinal mixed effects models . . . . .	35
	Exploratory analysis . . . . .	39
2.3	Experiment 1: Diatonic, Harmonic Minor, Jazz Minor . . . . .	39
2.3.1	Method . . . . .	40
	Participants . . . . .	40
	Stimulus . . . . .	40
	Procedure . . . . .	41
2.3.2	Results . . . . .	42
	Test for tonal hierarchies . . . . .	42
	Comparison of fit and stability ratings . . . . .	43
	Descriptive model . . . . .	45
	H2&3: Model 2.1 . . . . .	47
	Exploratory analysis . . . . .	49
2.3.3	Discussion . . . . .	52
2.4	Experiment 2: Less familiar scales . . . . .	52
2.4.1	Method . . . . .	53
	Participants . . . . .	54
2.4.2	Results . . . . .	54
	Test for tonal hierarchies . . . . .	54
	Comparison of fit and stability . . . . .	55
	Descriptive model . . . . .	55
	H2&3: Model 2.2 . . . . .	59
	Combined analysis: Model 2.3 . . . . .	61
	Exploratory analysis . . . . .	63
2.4.3	Discussion . . . . .	64
2.5	Conclusion . . . . .	65
<b>3</b>	<b>Experiments 1 &amp; 2 – Triads</b>	<b>67</b>
3.1	Introduction . . . . .	67
3.2	Overview of the Experiments and Models . . . . .	68

3.2.1	Hypotheses . . . . .	68
	Experiment 1 . . . . .	68
	Experiment 2 . . . . .	68
3.2.2	Analysis . . . . .	69
	Test for tonal hierarchies . . . . .	70
	Comparison of fit and stability ratings (H1) . . . . .	70
	Descriptive model . . . . .	71
	Testing H2 & 3 with Bayesian ordinal mixed effects models . . . . .	71
	Exploratory analyses . . . . .	74
3.3	Experiment 1: Diatonic, Harmonic Minor, Jazz Minor . . . . .	74
3.3.1	Method . . . . .	75
3.3.2	Results . . . . .	76
	Test for tonal hierarchies . . . . .	77
	Comparison of fit and stability ratings . . . . .	77
	Descriptive model . . . . .	79
	H2&3: Model 3.1 . . . . .	81
	Exploratory analysis . . . . .	86
3.3.3	Discussion . . . . .	88
3.4	Experiment 2: Less familiar scales . . . . .	89
3.4.1	Method . . . . .	89
	Participants . . . . .	90
3.4.2	Results . . . . .	90
	Test for tonal hierarchies . . . . .	90
	Comparison of fit and stability . . . . .	91
	Descriptive model . . . . .	91
	H2&3: Model 3.2 . . . . .	92
	Combined analysis: Model 3.3 . . . . .	96
	Exploratory analysis . . . . .	98
3.4.3	Discussion . . . . .	101
3.5	Conclusion . . . . .	102
<b>4</b>	<b>Experiment 3: Intrinsic stability of the triads of 22-TET</b>	<b>104</b>
4.1	Introduction . . . . .	104
4.2	22-TET . . . . .	105
4.3	Hypothesis . . . . .	108

4.4	Method . . . . .	108
4.4.1	Participants . . . . .	108
4.4.2	Stimuli and Procedure . . . . .	109
4.5	Analysis . . . . .	109
4.6	Results . . . . .	112
4.6.1	Descriptive analysis . . . . .	112
4.6.2	Hypothesis test . . . . .	113
4.6.3	Bayesian ordinal mixed effects regression model . . . . .	115
4.7	Discussion / Conclusion . . . . .	123
<b>5</b>	<b>Distributional Analysis of <math>n</math>-dimensional Feature Space for 7-note Scales in 22-TET</b>	<b>125</b>
5.1	Introduction . . . . .	125
5.2	Definitions . . . . .	126
5.3	Review . . . . .	127
5.3.1	Redundancy . . . . .	128
5.3.2	Coherence and Evenness . . . . .	131
5.3.3	$R$ -ad entropy . . . . .	133
5.3.4	Generator complexity . . . . .	133
5.3.5	Consonance . . . . .	134
5.3.6	Tetrachordality . . . . .	134
5.4	Analysis . . . . .	136
5.4.1	Reduction . . . . .	137
5.4.2	Cluster Analysis . . . . .	139
5.4.3	Exemplar Scales . . . . .	140
5.5	Conclusion . . . . .	145
<b>6</b>	<b>Experiments 4 &amp; 5: Stability of probe tones and triads in microtonal scales</b>	<b>147</b>
6.1	Introduction . . . . .	147
6.2	Hypotheses . . . . .	148
6.3	Method . . . . .	148
6.3.1	Participants . . . . .	148
	Experiment 4: Tones . . . . .	148
	Experiment 5: Triads . . . . .	148
6.3.2	Stimuli . . . . .	149



6.3.3	Procedure . . . . .	150
6.4	Analysis . . . . .	151
6.5	Results . . . . .	152
6.5.1	Test for tonal hierarchies . . . . .	152
6.5.2	Descriptive Model . . . . .	153
6.5.3	Confirmatory analysis . . . . .	163
6.5.4	Exploratory analyses . . . . .	167
6.6	Discussion and Conclusion . . . . .	174
<b>7</b>	<b>General Discussion and Conclusions</b>	<b>176</b>
	<b>Bibliography</b>	<b>180</b>
<b>A</b>	<b>Formal Specification of the Spectral Pitch Class Similarity Model</b>	<b>2</b>
<b>B</b>	<b>Experiments 1 &amp; 2 Tones</b>	<b>6</b>
B.1	Experiment 1 . . . . .	6
B.1.1	Pre-registered model . . . . .	6
B.1.2	Model 2.1 . . . . .	8
B.2	Experiment 2 . . . . .	14
B.2.1	Model 2.2 . . . . .	14
<b>C</b>	<b>Experiments 1 &amp; 2 Triads</b>	<b>19</b>
C.1	Experiment 1 . . . . .	19
C.1.1	Pre-registered model . . . . .	19
C.1.2	Model 3.1 . . . . .	20
C.2	Experiment 2 . . . . .	25
C.2.1	Model 3.2 . . . . .	25
C.2.2	Model 3.3 . . . . .	29
<b>D</b>	<b>Experiments 4 &amp; 5</b>	<b>33</b>
D.1	Experiment 4 Confirmatory . . . . .	33
D.2	Experiment 5 Confirmatory . . . . .	34
D.3	Experiment 4 Exploratory 1 . . . . .	36
D.4	Experiment 4 Exploratory 2 . . . . .	39
D.5	Experiment 4 Exploratory 3 . . . . .	40

## List of Tables

2.1	LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 1 . . . . .	43
2.2	Model 2.1 significant population-level effects . . . . .	47
2.3	LOOIC comparisons for Experiment 1 Bayesian ordinal mixed effects models . . . . .	49
2.4	LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 2 . . . . .	55
2.5	Model 2.2 Significant population-level effects . . . . .	60
2.6	Model 2.3 significant population-level effects . . . . .	62
2.7	LOOIC comparisons for Experiment 1 Bayesian ordinal mixed effects models. A significant negative $\Delta$ LOOIC supports the model shown in the first column; a significant positive $\Delta$ LOOIC supports the comparison model shown in the header. . . . .	63
3.1	LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 1 . . . . .	77
3.2	Model 3.1 significant population-level effects . . . . .	82
3.3	Evidence ratios for all significant effects of Model 3.1 . . . . .	84
3.4	LOOIC comparisons for Experiment 1 Bayesian ordinal mixed effects models . . . . .	86
3.5	LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 2 . . . . .	90
3.6	Model 3.2 significant population-level effects . . . . .	93
3.7	Evidence ratios for all significant effects of Model 3.2 . . . . .	94
3.8	LOOIC comparisons for Experiment 2 Bayesian ordinal mixed effects models . . . . .	96
3.9	Model 3.3 significant population-level effects . . . . .	97
3.10	Evidence ratios for all significant effects of Model 3.1 . . . . .	98

4.1	Model 4 selected population-level effects . . . . .	116
4.2	Perceived stability of the 16 most stable triads . . . . .	119
5.1	Exemplar scales associated with each successive cluster added. . . . .	143
5.2	Values of features for exemplar scales. . . . .	143
5.3	Z-scores of features for exemplar scales. . . . .	143
5.4	Values of features for exemplar scales. . . . .	146
6.1	LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 4 . . . . .	152
6.2	LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 5 . . . . .	153
6.3	Model Tone significant population-level effects . . . . .	164
6.4	Model Triad significant population-level effects . . . . .	165
6.5	LOOIC comparisons for Bayesian ordinal mixed effects models . . . . .	166
6.6	Exploratory Model 6.1 population-level effects . . . . .	168
6.7	Evidence ratios for all significant effects of Exploratory Model 6.1 . . . . .	169
6.8	Exploratory Model 6.3 population-level effects . . . . .	173
B.1	Pre-registered probe tones model significant population-level Effects	7
B.2	All population-level effects for Model 2.1, which was summarized in Table 2.2 . . . . .	8
B.3	Model 2.1 but without Prevalence significant population-level effects	11
B.4	Model 2.1 but with ScaleTone instead of SPCS significant population- level effects . . . . .	12
B.5	Model 2.1 but without Prevalence and with ScaleTone instead of SPCS significant population-level effects . . . . .	13
B.6	All population-level effects for Model 2.2, which was summarized in Table 2.5 . . . . .	14
B.7	All population-level effects for Model 2.3, which was summarized in Table 2.6 . . . . .	17
C.1	Pre-registered probe triads model significant population-level effects	20
C.2	All population-level effects for Model 3.1, which was summarized in Table 3.2 . . . . .	21
C.3	Model 3.1 but without Prevalence significant population-level effects	23

C.4	Model 3.1 but with ScaleTone instead of SPCS significant population-level effects . . . . .	24
C.5	Model 3.1 but without Prevalence and with ScaleTone instead of SPCS significant population-level effects . . . . .	25
C.6	All population-level effects for Model 3.2, which was summarized in Table 3.6 . . . . .	26
C.7	Model 3.2 but with ScaleTone instead of SPCS significant population-level effects . . . . .	29
C.8	All population-level effects for Model 3.3, which was summarized in Table 3.9 . . . . .	29
C.9	Model 3.3 but with ScaleTone instead of SPCS significant population-level effects . . . . .	32
D.1	All population-level effects for Model Tone, which was summarized in Table 6.3 . . . . .	33
D.2	All population-level effects for Model Triad, which was summarized in Table 6.4 . . . . .	34
D.3	All population-level effects for Exploratory Model 6.1, which was summarized in Table 6.6 . . . . .	36
D.4	All population-level effects for Exploratory Model 6.2 . . . . .	39
D.5	All population-level effects for Exploratory Model 6.1, which was summarized in Table 6.8 . . . . .	40

# List of Figures

2.1	Average fit vs stability ratings for probe tones after the diatonic context. Scale-tones are numbered as RPCs. Error bars are from 95% confidence intervals obtained from 1000 bootstrapped samples. . . .	44
2.2	Average fit vs stability ratings for probe tones after the harmonic minor context . . . . .	44
2.3	Average fit vs stability ratings for probe tones after the jazz minor context . . . . .	44
2.4	Average ratings for probe tones after the diatonic context compared to SPCS predictions . . . . .	46
2.5	Average ratings for probe tones after the harmonic minor context compared to SPCS predictions . . . . .	46
2.6	Average ratings for probe tones after the jazz minor context compared to SPCS predictions . . . . .	46
2.7	<i>ppcheck</i> – posterior predictive check – grouped by scale-probe combination. Each plot shows the number of each of the 7 Likert scale ratings (from ‘very unstable’ or ‘very bad fit’ on the left to ‘very stable’ or ‘very good fit’ on the right) for each probe after each scale observed ‘ <i>y</i> ’ and predicted by the model ‘ <i>yrep</i> ’. The first number in the title for each plot refers to the scale, where ‘1’ is the diatonic, ‘2’ is the harmonic minor and ‘3’ is the jazz minor; the second number in each pair refers to the RPC of the probe. ‘ <i>y</i> ’ represents the data and ‘ <i>yrep</i> ’ represents the distribution of predictions obtained from random samples of parameter values from the posterior predictive distribution for Model 2.1. . . . .	50
2.9	Average ratings for probe tones after the double harmonic context compared to SPCS predictions . . . . .	57
2.10	Average ratings for probe tones after the pentatonic context compared to SPCS predictions . . . . .	57

2.8	Average ratings for probe tones after the harmonic major context compared to SPCS predictions . . . . .	57
2.11	Average ratings for probe tones after the blues context compared to SPCS predictions . . . . .	58
2.12	Average ratings for probe tones after the hexatonic context compared to SPCS predictions . . . . .	58
2.13	Average ratings for probe tones after the octatonic context compared to SPCS predictions . . . . .	58
3.1	Average fit vs stability ratings for probe triads after the diatonic context. Triads are labelled by the Roman numeral for the scale-tone of their root. . . . .	78
3.2	Average fit vs stability ratings for probe triads after the harmonic minor context . . . . .	78
3.3	Average fit vs stability ratings for probe triads after the jazz minor context . . . . .	78
3.4	Average ratings for probe triads after the diatonic context compared to predictions due to SPCS and Triad . . . . .	80
3.5	Average ratings for probe triads after the harmonic minor context compared to predictions due to SPCS and Triad . . . . .	80
3.6	Average ratings for probe triads after the jazz minor context compared to predictions due to SPCS and Triad . . . . .	80
3.7	Conditional effect of MusSoph:Triad for Model 3.1 . . . . .	84
3.8	Conditional effect of Triad for Model 3.1 . . . . .	85
3.9	Average ratings for probe tones after the harmonic major context compared to predictions due to SPCS and Triad . . . . .	91
3.10	Average ratings for probe tones after the double harmonic context compared to predictions due to SPCS and Triad . . . . .	92
3.11	Conditional effect of Triad for Model 3.2 . . . . .	94
3.12	Conditional effect of MusSoph:Triad for Model 3.2 . . . . .	95
3.13	Conditional effect of MusSoph:Triad for Model 3.3 . . . . .	99
3.14	Conditional effect of Triad for Model 3.3 . . . . .	99
3.15	Conditional effect of TrialNumber:Triad for Model 3.3 . . . . .	100
3.16	Conditional effect of InContTrialNo:Triad for Model 3.3 . . . . .	100

4.1	A visualisation of 12-TETs and 22-TETs approximation of 11-integer-limit frequency ratios . . . . .	107
4.2	Average stability ratings for all triads of 22-TET . . . . .	114
4.3	The 15 most stable triads of 22-TET, notated using Ups and Downs . . . . .	120
4.4	22-TET Ups and Downs notation guide, from <i>Tibia 22edo ups and downs guide 1</i> by Kite Giedraitis, 2016. Downloaded from <a href="https://en.xen.wiki/w/File:Tibia_22edo_ups_and_downs_guide_1.png">https://en.xen.wiki/w/File:Tibia_22edo_ups_and_downs_guide_1.png</a> . Used with permission. . . . .	121
5.1	3D clustering view 1 . . . . .	141
5.2	3D clustering view 2 . . . . .	142
6.1	Average ratings for probe tones after a context of Scale 1, compared to SPCS predictions – scale-tones only, numbered as RPCs. . . . .	154
6.2	Average ratings for probe tones after a context of Scale 1, compared to SPCS predictions – all RPCs, with scale-tones numbered. . . . .	154
6.3	Average ratings for probe triads after a context of Scale 1, compared to SPCS predictions. Triads roots are labelled by Roman numeral. . . . .	154
6.4	Average ratings for probe tones after a context of Scale 2, compared to SPCS predictions – scale-tones only. . . . .	155
6.5	Average ratings for probe tones after a context of Scale 2, compared to SPCS predictions – all RPCs. . . . .	155
6.6	Average ratings for probe triads after a context of Scale 2, compared to SPCS predictions. Triads roots are labelled by Roman numeral. . . . .	155
6.7	Average ratings for probe tones after a context of Scale 3, compared to SPCS predictions – scale-tones only. . . . .	156
6.8	Average ratings for probe tones after a context of Scale 3, compared to SPCS predictions – all RPCs. . . . .	156
6.9	Average ratings for probe triads after a context of Scale 3, compared to SPCS predictions. Triads roots are labelled by RPC. . . . .	156
6.10	Average ratings for probe tones after a context of Scale 4, compared to SPCS predictions – scale-tones only. . . . .	157
6.11	Average ratings for probe tones after a context of Scale 4, compared to SPCS predictions – all RPCs. . . . .	157
6.12	Average ratings for probe triads after a context of Scale 4, compared to SPCS predictions. Triads roots are labelled by Roman numeral. . . . .	157

6.13	Average ratings for probe tones after a context of Scale 5, compared to SPCS predictions – scale-tones only. . . . .	158
6.14	Average ratings for probe tones after a context of Scale 5, compared to SPCS predictions – all RPCs. . . . .	158
6.15	Average ratings for probe triads after a context of Scale 5, compared to SPCS predictions. Triad roots are labelled by RPC. . . . .	158
6.16	Average ratings for probe tones after a context of Scale 6, compared to SPCS predictions – scale-tones only. . . . .	159
6.17	Average ratings for probe tones after a context of Scale 6, compared to SPCS predictions – all RPCs. . . . .	159
6.18	Average ratings for probe triads after a context of Scale 6, compared to SPCS predictions. Triad roots are labelled by RPC. . . . .	159
6.19	Average ratings for probe tones after a context of Scale 7, compared to SPCS predictions – scale-tones only. . . . .	160
6.20	Average ratings for probe tones after a context of Scale 7, compared to SPCS predictions – all RPCs. . . . .	160
6.21	Average ratings for probe triads after a context of Scale 7, compared to SPCS predictions. Triad roots are labelled by RPC. . . . .	160
6.22	Average ratings for probe tones after a context of Scale 8, compared to SPCS predictions – scale-tones only. . . . .	161
6.23	Average ratings for probe tones after a context of Scale 8, compared to SPCS predictions – all RPCs. . . . .	161
6.24	Average ratings for probe triads after a context of Scale 8, compared to SPCS predictions. Triad roots are labelled by RPC. . . . .	161
B.1	Conditional effect of Height:TriadNo for Model 2.1 . . . . .	10
B.2	Conditional effect of RelHeight for Model 2.2 . . . . .	10
B.3	Conditional effect of RelHeight for Model 2.2 . . . . .	16
B.4	Conditional effect of Height:ScaleSize for Model 2.2 . . . . .	16
C.1	Conditional effect of BlockOrder:Triad for Model 3.2 . . . . .	28



# List of Abbreviations

SPCS	Spectral Pitch Class Similarity
ET	Equal Temperament
12-TET	12-Tone Equal Temperament
22-TET	22-Tone Equal Temperament
RPC	Relative Pitch Class
PSIS	Pareto Smoothed Importance Sampling
LOOIC	Leave-One-Out cross validation Information Criterion
SE	Standard Error
JI	Just Intonation
MOS	Moment of Symmetry
ME	Maximally Even
DE	Distributionally Even
WF	Well-Formed
PWF	Pairwise Well-Formed
SGC	Scalar Graham Complexity
VIF	Variance Inflation Factor
PCA	Principal Components Analysis

## *Abstract*

Music is ubiquitous across all human cultures. It is hypothesised that the development of music and of language in human evolution is linked (Wallin et al., 2001), and music, in addition to language, is known to be communicative. One way music – particularly music employing the widely used system of tonality – communicates is through tension and resolution, or stability and instability, where instability is the need to resolve and stability its destination. Most tonal-harmonic music today exists in a Western tuning system and experimental research into the perception of harmonic tonality is conducted almost entirely in 12-TET. This project is the first empirical study of the cognition of harmonic tonality in microtonal scales. Through the employment of novel scales in an unfamiliar tuning system, effects of familiarity are weakened, allowing a more focussed investigation of other effects. Particularly, bottom-up models for the cognition of harmonic tonality are allowed a more careful investigation, providing valuable insight into the cognition of music otherwise beyond reach. This research also provides valuable information for hopeful composers of novel music in shaping their music to elicit a desired response, thus enabling expansion of the palette of possible musical expression. This project utilizes a common experimental paradigm for research into the cognition of tonality: participants are first played context-setting stimuli, after which a probe tone or chord is sounded and they are asked to rate how well the probe tone “fits” the context, or how stable it is given the context. A psychoacoustic feature – spectral pitch class similarity – is used to predict the perceived stability of pitch classes and triads of not only familiar scales (Experiment 1), but unfamiliar (Experiment 2), and novel scales (Experiments 3-5), where models of long-term statistical learning are available only for familiar scales. Through a series of 5 experiments the perceived stability of tones and triads in novel, microtonal scales is predicted, demonstrating the usefulness of our psychoacoustic model.

# Chapter 1

## Background

### 1.1 Introduction

#### 1.1.1 Tonality and Scale

The music of many cultures is considered to be *tonal* (Krumhansl, 2001), wherein there exists a hierarchy of perceived stability of pitch classes (sets of pitches of the same *chroma* – chromatic pitch class name – that may be any number of octaves apart) or chords (sets of two or more pitch classes) within a scale. The most stable pitch class is referred to as the *tonic*, which is said to function psychologically as a reference point with which other pitch classes possess a certain relationship (Dahlhaus et al., 1980; Hyer, 2002).

These relationships may involve the formation of a hierarchy of stability which, if exploited, may induce feelings of tension and resolution in the listener, allowing the communication of emotional meaning. A typical example of tension and resolution in traditional Western tonality is the tendency for the less stable leading-tone, the pitch class immediately below the tonic, to resolve up to the tonic, the most stable pitch class (Krumhansl, 1990). Tonal music that places a particular importance on harmony, as in most Western music of the common practice period (from about 1600 to 1900) and much contemporary Western music is described as ‘tonal-harmonic’ music (Krumhansl, 1990), wherein the hierarchical units comprise triads more so than individual pitch classes (R. Parncutt, 2011). In such music the tonic exists as a chord as well as as a tone, the most stable chord also carrying the label ‘tonic’. Western music involves the use of ‘tonic triads’. We suggest that both pitch classes and chords are related hierarchically to the tonic pitch class and chord.

In this project we consider mostly the tertian triads (meaning three-note chords constructed ‘by thirds’) made entirely of scale-tones (pitch classes that belong to the

scale), which result from taking, after choosing a scale-tone, the scale-tones two and four scale-steps above (or below). Some tertian triads containing non-scale-tones are also tested in the final experiment, discussed in Chapter 6.

For example, the ordered set of tertian triads of the A harmonic minor scale – A B C D E F G $\sharp$  – are A-C-E, B-D-F, C-E-G $\sharp$ , D-F-A, E-G $\sharp$ -B, F-A-C and G $\sharp$ -B-D.

These triads are written in *root position*, where above the triad's *root* is the pitch two scale-tones above – in music theory, this is denoted the triad's *third* – and the pitch four scale-steps above – the triad's *fifth*. Triads may also occur in inversion, comprising the same pitch classes, but in a different pitch height order. For example, the triad C4-E4-A4 (the '4' after the pitch classes specifies that their pitch is in the fourth octave, where C4 is the lowest pitch of the three) is still considered an A minor triad, but rather than root position it is said to be in *first inversion*, where instead of a lowest pitch of the root 'A', the third 'C' is the lowest pitch. Similarly, E-A-C is labelled the *second inversion* tonic triad, where the fifth 'E' is the lowest pitch.

Tertian triads are typically most commonly of four types depending on the number of semitones above the root the third and fifth lie. Triads with a *perfect fifth* (seven semitones above the root) are labelled *major* or *minor*, depending on whether the third is major (four semitones above the root) or minor (three semitones above the root). The triad with an *augmented fifth* (eight semitones above the root), and a major third is called an *augmented triad*, and the triad with a *diminished fifth* (six semitones above the root) and a minor third is called a *diminished triad*. Triads containing non-scale-tones (pitch classes that are not members of the scale) are, here, labelled *chromatic*. Non-tertian triads also exist, but are less commonly used in tonal-harmonic music, and will not be discussed in this dissertation.

Throughout this dissertation (specifically in Chapters 1-3 and 6), triads are labelled by the Roman numeral for the scale-tone of their root, this being standard notation practice in Western music. For example, the triad rooted on the third note of the scale is labelled 'iii', for the diatonic scale, the triad rooted on the fifth note of the scale is labelled 'V'. Minor triads are labelled with lower case roman numerals and major triads with uppercase. Diminished triads are labelled with lower case roman numerals followed by '°', and augmented triads with upper case roman numerals followed by '+'. For example, the scale-tone tertian triads of the harmonic minor scale can be labelled 'i', 'ii°', 'III+', 'iv', 'V', 'VI', and 'vii°', which tells us that triads rooted on the first (the tonic) and fourth notes of the scale are minor, the triads on the fifth and sixth notes of the scale are major, the triads on the second and

seventh (leading-tone) of the scale are diminished, and the triad on the third note of the scale is augmented.

Three important scales in Western music are the diatonic or major (e.g., C, D, E, F, G, A, B), the harmonic minor (e.g., A, B, C, D, E, F, G $\sharp$ ), and ascending melodic minor (jazz minor) (e.g., A, B, C, D, E, F $\sharp$ , G $\sharp$ ). They are important because they form the basis of most elementary music theory text books and are commonly used in music. In traditional Western harmonic tonality, only major or minor triads are used as tonic chords. Scales are used for melody and harmony, and these scales typically have unequally sized steps. The major mode of the diatonic scale is widely used to reinforce the major triad as tonic. The minor triad is often reinforced as a tonic by use of the harmonic minor scale, though the natural and melodic minor scales are also often employed. The natural minor scale comprises the same pitch classes as its *relative* major scale, but starting on a different pitch class, which in this example is used as a tonic. For example, the relative natural minor scale to C major (introduced above – C, D, E, F, G, A, B) is the A natural minor scale – A, B, C, D, E, F, G. The major scale and the natural minor scale are two *modes* (rotations) of the diatonic scale. The *melodic minor* scale is equivalent to the jazz minor scale when ascending, and to the natural minor scale when descending. Progressions of chords from the pitches of these scales lead to resolution on the tonic triad.

### 1.1.2 Thesis Summary

A series of five experiments were run in order to investigate whether the perceived stabilities of pitch classes and tertian triads in a scale follow directly from the structure of that scale and of the triads. In other words, with no additional cues such as temporal ordering or prevalence, are some pitch classes and triads more likely to be perceived as stable and some more likely to be perceived as unstable? The first two experiments also investigate the possible equivalence between ratings of fit and stability. The first experiment tests the perceived fits and stabilities of tones and triads in the three common scales – the diatonic, harmonic minor, and jazz minor (melodic minor ascending). In addition to Spectral Pitch Class Similarity (SPCS), Prevalence, a measure of the frequency of occurrence of tones and triads in corpora of Western rock and pop music, was found to be able to predict ratings of fit and stability. Experiment 2 then tests for the perceived fits and stabilities of tone and triads from 6 different scales of 12-TET (12-tone equal temperament) for which no

prevalence data are available, 4 of which are much less familiar. Since these scales though, however unfamiliar, are not completely foreign to Western music, we turn to microtonal music for novel scales in which familiarity may be further reduced. Experiment 3 then tests the intrinsic perceived stability of all triads in 22-TET (i.e., due to the triads themselves without any contextual effects). Following Experiment 3 a cluster analysis is run in order to find exemplar scales representing all possible 7-note scales of 22-TET: Given a selection of independent scale features, including the intrinsic stability of triads found within the scales, a set of scales most representative of the entire set of possibilities are found such that they can be used as context scale in the final pair of experiments. Experiments 4 and 5 are such a pair, testing the perceived stabilities of pitch classes and triads, respectively, of 8 seven-note scales of 22-TET. The probe tone / triad experimental paradigm is employed in all experiments.

## 1.2 Background

### 1.2.1 Sound, spectrum, pitch and tuning

Sound is caused by, and permeates through, vibrating objects. Musical tones comprise frequencies of vibration at a *fundamental*, lowest frequency, in addition to vibrations of higher frequencies called *overtones*. All of these frequencies are *harmonics* or *partials* of the tone, and their relative intensities comprise the frequency *spectrum* of the tone. An ideal string, for example, vibrates as a whole at a fundamental frequency, in two halves at twice the fundamental frequency, in three thirds at three times the fundamental frequency, and so on. The harmonics of a perfect string make up the *harmonic series*. The spectrum of many musical instruments resemble, but do not exactly match the harmonic series. The interval between the first overtone, or second harmonic of a fundamental can be described by a frequency ratio of 2/1. Tones separated by this interval – the simplest possible interval other than the 1/1 interval of a *unison* – are considered to have the same chroma and are given the same pitch class name – ‘C’ or ‘A’, for example (where ‘C’ may refer to C1, C2, or C4, etc.). Musical scales of most cultures repeat at this interval. Since Western scales typically had, and have, 7 pitch classes, this interval is called an *octave* (referencing its identity as the 8th note of a most scales). The mechanism by which tones separated by an octave are perceived as equivalent is called *octave equivalence*, and has been

observed experimentally (Hoeschele et al., 2012). Then, each doubling of frequency corresponds to the same linear difference in pitch. Perceived pitch is essentially logarithmic to frequency, though pitch perception is not so simple, also depending on other factors (De Cheveigne, 2005; Plack et al., 2006). In this thesis 'pitch' refers to *musical pitch*, which is proportional to  $\log f$ , rather than perceived pitch.

Western tonal-harmonic music is today tuned most commonly, theoretically, to 12-TET (12-tone equal temperament), in which the octave is divided into 12 pitch classes, equally distant from each other in pitch<sup>1</sup>. 12-TET is one of an infinite number of possible tuning systems, some of which may be more suited to harmonic tonality than others. A *tuning system* is a set of musical pitches from which subsets (*scales*) are employed for melody and harmony in musical composition and performance.

An intriguing possibility is to produce tonal-harmonic music in a novel (microtonal) tuning, employing novel scales (novel collections of musical pitches) (Erlich, 1998). Such a system would expand the possibility for musical expression for a composer loosely akin to a painter discovering how to employ a new set of colours. Composition in novel musical tunings is itself not novel, with notable work by Vicentino (Vicentino, 1996), Blackwood (Blackwood, 1980), Carlos (Carlos, 2000), Hába (Hába, 1927), Carrillo (Madrid, 2015), Novaro, Johnston (Johnston, 1984) Dean (Dean et al., 2008), Miller, Walker, Sethares (Sethares, 2005), Branca (Branca, 1993), Chowning (Chowning, 2012), Xenakis (Dean et al., 2008), Erlich ("22edo", 2020), Andrew Milne, and many others. I have also produced such work ("22edo", 2020; Hearne, n.d.). Harmonic tonality involves both consonance and affinity. *Consonance* (and *dissonance*, its antonym), describes the perceived agreement of simultaneous sonorities (Malmberg, 1918) and *affinity* describes the perceived agreement of successive sonorities (Milne et al., 2016; R. Parncutt & Hair, 2011; Terhardt, 1984). In this project, a series of experiments will be developed in order to assess to what extent the employment and perception of harmonic tonality in a novel microtonal tuning system is possible. More directly we test to see if we can predict the stability of tones and triads in microtonal scales. We also aim to consider to what extent the structure, consonance and affinity of the chords and pitch collection used affect the scope of

---

<sup>1</sup>Although only in computer music can 12-TET be exactly achieved, and only on keyboards is it used with almost exact tuning. Microtonal inflections are a reality in Western music, whether stylistic (Fabian et al., 2014) – "blue" notes, for example – or simply due to inaccuracies in the tuning of instruments. 12-TET, then, is perhaps better understood as a theoretical scaffolding from which tuning is referenced. The same could be true of other ETs and tuning systems

the system for the communication of emotional meaning, though exploration of this is mostly left for future research.

While we may speculate at this stage that it may be possible to produce tonal-harmonic music in any tuning system, for parsimony, I choose a single tuning system in which to conduct Experiments 3-5. As equal temperaments are the most simply structured tuning systems and allow free transposition, an equal temperament is chosen for the tuning system used in these experiments. As there exists no empirical research yet into the effect of a novel, microtonal context on the perception of pitches and triads, no assumptions will be made of their perceived stability, and thus all possible triads within the equally tuned system will be considered. More finely grained  $n$ -TETs (i.e. when  $n$  is a large number) would, therefore, be impractical. Most psychoacoustic models of consonance rate intervals that approximate low integer ratios of frequency as more consonant than those that do not (Large et al., 2016). After the familiar 12-TET, the smallest equal temperament including close approximations to the same low integer ratios as 12-TET, as well as the next lowest integer ratios, is 22-TET. Any equal temperament (larger than 6-TET) includes small intervals, which are universally considered to be dissonant (Stolzenburg, 2015). Therefore 22-TET (like 12-TET) includes possibilities for both dissonance and consonance. A tuning of 22-TET includes many different possible scales and chords that may be transposed freely and may exhibit varied consonance and affinity, while still being just small enough to logistically allow a complete exploration of the perception of its triads. For this reason, I will employ the use of 22-TET in the final pair of experiments in this project.

Though tonal harmonic music is many-dimensional in its construction, harmony is hypothesized to be its most perceptually salient organisational feature. Supporting this hypothesis, Krumhansl et al. (1987) interpret from the results of their experiments that 'the underlying organization of the sequences was perceived as largely unaffected by the rhythmic and melodic variations of the excerpts' (Krumhansl et al., 1987, p. 74). For this reason, these features of musical organisation will not be stressed in this project. The effect of beat, rhythm, melody and phrase is also beyond the scope of this project; however, Bigand (1997) observes that these properties of music have a smaller effect upon the perception of stability in harmonic tonality than harmony. In summary, the effects of the placement of chords within the phrase and beat structure of the music, though understood as a salient organisational feature in the perception and cognition of music, are beyond the scope of this project



and are not tested here.

In this dissertation, though all *modes* (rotations) are tested in the experiments, each scale is expressed in a single *exemplar mode* (rotation) that, when possible, is the most commonly used mode of that scale – i.e. the lowest note of the mode is most commonly used as a tonic – when that is clear. We define the *relative pitch class* or *RPC* of a pitch class as its distance in degrees of an ET above the pitch class of the first degree of the exemplar mode of the scale, which may function psychologically as the tonic. Tertian triads are defined after the RPC of their roots. For the experiments in which stimuli are tuned to 12-TET, the relative pitch class is represented by an integer from 0 to 11, where a unit corresponds to a single degree of 12-TET – a (12-TET) semitone. When 22-TET is instead employed, relative pitch classes are represented by the integers 0 to 21, a unit corresponding again to a single degree of the equal temperament, though this time the step being much smaller ( 55 cents – 100ths of a 12-TET semitone).

### 1.2.2 Consonance and Affinity

Consonance and affinity are widely considered to depend on both psychoacoustic and cultural processes Parncutt and Hair (2011). Consonance depends upon harmonicity – the degree to which simultaneous sonorities together resemble a harmonic complex tone, and its opposite, dissonance, upon roughness due to beating in the ear canal, which are both psychoacoustic processes. Consonance and dissonance are also affected by familiarity, which is often modelled by prevalence in a relevant musical corpus. Affinity is considered to depend again upon familiarity as a cultural component, along with the psychoacoustic processes pitch proximity and pitch commonality (R. Parncutt & Hair, 2011), which may be modelled by pitch similarity. Judgements of affinity are also affected by consonance, where consonant sonorities are judged to fit together better than dissonant sonorities (Milne et al., 2016).

Cultural processes are often termed ‘top-down’, and are understood to invoke higher cognitive processing such as statistical learning. Psychoacoustic processes are often termed ‘bottom-up’. Bottom-up models of consonance and affinity do not rely upon prior knowledge of statistical regularities in music (Milne, Laney, et al., 2015). A bottom-up model may allow the prediction of the distribution of notes in a novel tunings system that leads to a desired result, such as a hierarchy of differing

stabilities and a strong tonic chord. Understanding that the cognition of tonality depends on both bottom-up and top-down processes, my research will explore possible bottom-up contributions to the cognition of tonality through the employment of unfamiliar scales and a novel tuning system.

Nomenclature considering the perception of harmony is far from standardized: Within the literature many different words carry many different meanings. Bottom-up aspects of consonance are often referred to as ‘sensory consonance’ (Izumi, 2000; Regnault et al., 2001; Schellenberg & Trainor, 1996; Terhardt, 1984; Trainor et al., 2002) or ‘sensory dissonance’, (Bigand et al., 1996; Mashinter, 2006). Sensory consonance is typically considered to be consonance for isolated musical intervals without musical context (Vos, 1982) and depends upon psychoacoustic properties. Top-down aspects of consonance are often referred to as ‘musical consonance’ (Plack, 2010), depending upon learned associations due to familiarity with a corpus of tonal music. Use of the single word ‘consonance’ varies considerably however, referring sometimes to sensory consonance, sometimes to musical consonance and other times to a combination of the two. While some authors describe musical consonance, ‘tonal stability’ (Large, 2010; Large et al., 2016; Lerud et al., 2014), ‘harmonic stability’ (McDaniel & Williams, 2012), ‘musical stability’ (Bigand, 1997), or ‘tonal dissonance’ (Johnson-Laird et al., 2012). Where some authors refer to ‘sensory consonance’, others refer to ‘tonal consonance’ (Huron, 1991, 1994; Kameoka & Kuriyagawa, 1969a, 1969b; Plomp & Levelt, 1965), ‘the acoustic component of consonance’ or ‘acoustic consonance’ (Hutchinson & Knopoff, 1978) and local consonance (Sethares, 1993), and, confusingly, even ‘musical consonance’ (van de Geer et al., 1962). In an effort to disentangle terminology, McDaniel and Williams (2012) presents a case for the separation of these ideas into the terms ‘consonance’, referring to sensory components of consonance, i.e. roughness and harmonicity, and ‘harmonic stability’, which, ‘in Western music, deals with a sonority’s location within a tonal pitch space’ (McDaniel & Williams, 2012, p. 11).

Parncutt and Hair (2011) discuss consonance as dependent upon temporal smoothness (the opposite of roughness), spectral harmonicity and cultural familiarity. They then ‘compare and contrast dichotomies that overlap or interact with the [consonance/dissonance] concept such as tense/relaxed, primary/subordinate, centric/acentric, diatonic/chromatic, stable/unstable, close/distant, similar/different, rough/smooth, fused/segregated, related/unrelated, familiar/unfamiliar,

implied/realized, tonal/atonal', suggesting that 'our perception of these dichotomous pairs often intensifies, parallels or stands in for our perception of [consonance/dissonance]' (R. Parncutt & Hair, 2011, p. 119). Affinity, may be described by 'goodness-of-fit' (Krumhansl, 1990; Milne et al., 2016) 'musical tension' (Bigand & Parncutt, 1999; Bigand et al., 1996), 'tonal tension' (Lerdahl, 1996; Lerdahl & Krumhansl, 2007), 'harmonic tension' and 'melodic attraction' (Vega, 2003), 'musical stability' (Bigand, 1997) or 'tonal stability' (Large et al., 2016).

My research will test and refer simply to perceived 'stability', which is presumed to be affected by both consonance and affinity, which may depend on any combination of the above ideas. Where Cook et al. (2004) questionably assumes operational equivalence between stability and 'harmoniousness, sonority, tonalness, tonality, etc.' (Cook et al., 2004, p. 494) I will not assume those equivalences.

Recent studies sought to tease apart the effects of the different theories of consonance. Cousineau et al. (2012) tested responses to stimuli in individuals with cognitive amusia, finding no preference for consonance over dissonant intervals, or for harmonic over inharmonic tones. Though the participants were unable to distinguish between harmonic and inharmonic tones, they showed normal preference and discrimination for stimuli with and without beating. These results suggest that harmonicity is more likely than sensory dissonance to underlie consonance. McDermott et al. (2016) similarly finds no preference for consonant chords in the responses of native Amazonian tribes completely unfamiliar with Western music, while finding such preferences in participants from Bolivia and the US that are familiar with Western music, supporting the top-down aspect to perceived consonance, wherein preferences are influenced by enculturation / long-term statistical learning. Parncutt et al. (2019) found that a consonance model of roughness, harmonicity and familiarity was able to predict the prevalences of trichords in a corpus of Western polyphonic vocal music. Smit et al. (2019) found that perceived pleasantness and happiness of microtonal triads could be predicted by a model of roughness, harmonicity, spectral entropy, average pitch height and proximity to pitches of 12-TET. Clearly there are many contributions to perceived consonance, encompassing both innate and learned processes. In this thesis we do not seek to model consonance, representing it instead with categories of triad type, e.g., major, minor, diminished or augmented. The stability of triads, of which consonance is one aspect, is modelled instead.

### 1.2.3 Probe Tone Experiments

The probe tone paradigm, employed in a series of experiments on the cognition of tonality including those in this dissertation, was introduced in Krumhansl and Shepard (1979). In their study participants were first played an ascending or descending C major scale (beginning on C and concluding on B or D) as a context setting stimulus, followed by, as a probe tone, any pitch of the chromatic scale, from middle C to the C an octave above. The descending scale was presented in the octave above the range of probes, and the ascending scale in the octave below. Participants were asked to rate how well the probe tone completed the scale. The two Cs received the highest rating, and the pitch classes of the C major scale received higher ratings than the chromatic pitch classes (the non-scale-tones). However, the pitch height proximity of the probe to the scale influenced participants' ratings; they typically gave a higher rating to pitches closer in pitch height to the final pitch(es) of the scale. Though a flute stop on a Farfisa electronic organ was used as an approximation to a sine wave for stimulus, the authors reasoned that there may still be an influence by 'some perception of overlap between the harmonics present in the test tone and the frequencies of the preceding context tones' (Krumhansl & Shepard, 1979, p. 589). In an effort to address this, a second experiment employed computer generated sine tones for the stimulus. The results followed the same trends but had an even stronger effect of pitch height on ratings.

In Krumhansl and Kessler (1982) the procedure was applied again but in order to minimize the effect of pitch height Shepard tones<sup>2</sup> were employed for the stimulus instead of sine tones. Participants judged how well a probe tone fit the context stimulus, which was of three types: either an ascending major or harmonic minor scale (this time with the tonic pitch class repeated); a single major, minor, diminished or dominant seventh chord, each played three times; or a conventional three chord cadence in a major or minor key (IV-V-I, II-V-I, or VI-V-I). These stimuli were designed in order to induce a specific key. Of these, the repeated major and minor chords and the cadences were found to produce highly correlated results, and the remaining context types were not included in the principal analysis. Due perhaps mostly to the use of Shepard tones, the effects of pitch height were decreased, as compared to the 1979 study, though for scalic contexts an effect of proximity to the

---

<sup>2</sup>Octave complex tones with clear pitch chroma but ambiguous pitch height (Krumhansl & Kessler, 1982). Shepard tones do not have a pitch height, and so pitch range is meaningless in stimuli using such tones.

last note of context was still observed. The highest ratings were given to the roots of the final chord; the next highest ratings were given to the other pitch classes of this (tonic) triad; the next highest to the remaining pitch classes of the major and minor scales whose tonics correspond to these triads; and the lowest to the chromatic pitch classes (non-scale-tones).<sup>3</sup>

Krumhansl posits that the cognition of harmonic tonality is due to statistical learning, supporting Meyer's view (Meyer, 2008) that 'through experience with music, listeners have abstracted and internalized certain probabilistic regularities underlying the musical tradition' Krumhansl (2001, p. 75). Krumhansl suggests that through enculturation to Western music, humans are able to internalize the Western tonal hierarchy. Her results seemed to support this view: the prevalence of pitches within a key in Western common-practice music provides a reasonably accurate model of her data (Krumhansl, 2001). Given unfamiliar contexts, however, such as microtonal music (Leung & Dean, 2018; Loui & Wessel, 2008; Loui et al., 2010) or, to Western listeners, the music of non-Western cultures (Castellano et al., 1984; Kessler et al., 1984; Lantz et al., 2014), several studies have demonstrated that perceived hierarchy can be formed and/or altered rapidly. This suggests that the hierarchy-forming statistical learning may also be due to a less long-term process or processes, though Castellano et al. (1984) suggests that internalization of a tonal hierarchy requires more extensive exposure than provided in the experiment.

Krumhansl aimed to remove any possible effect of overlap of harmonics between context and probe in her early probe tone experiments, through the use of sine tones. Later, Shepard tones were used to minimize effects of pitch height; however, real music makes minimal use of such tones, which cannot be created through any natural physical process, but must be artificially synthesized (Milne, Laney, et al., 2015). Further, it has been shown that the response of inner hair cells in the basilar membrane is nonlinear and nonlinearities transform a pure tone into a harmonic complex tone (Pickles, 1988). Studies have also demonstrated that due to non-linear distortion in the auditory system the human brainstem response to two pure tones additionally includes combination and difference tones (Chertoff & Hecox, 1990; Chertoff et al., 1992; Elsisy & Krishnan, 2008; Galbraith, 1994; Greenberg et al., 1987;

---

<sup>3</sup>Krumhansl and Kessler wrote that the average rating of all non-tonic scale-tones was higher than that of the non-scale-tones. They specified that the harmonic form or the minor scale was used. The raised seventh however – the seventh of the harmonic minor scale – received lower ratings than the natural seventh, so the way we describe the results is true for the natural minor scale, and not for the harmonic minor scale.

Krishnan, 1999; Pandya & Krishnan, 2004; Rickman et al., 1991). Lee et al. (2009) provides a succinct review of such studies. A Shepard tone includes pure tones with frequencies  $f$  and  $2f$ . One resulting combination tone, the summation tone, has frequency  $f + 2f = 3f$  resulting already in a brainstem response characteristic of a harmonic complex tone. In summary, nonlinear distortion – such as is exhibited in the auditory brainstem – will transform a Shepard tone into a full harmonic complex tone. We assume, therefore, that although presented with Shepard tones, participants would complete the cognitive task of rating the fit of the probe tone in consideration to harmonic complex tones including non-octave harmonics, and that any effect of overlap of harmonics between context and probe within the auditory system cannot be avoided.

An alternative explanation for the cognition of tonality depends upon this overlap that Krumhansl sought to avoid. Building on Parncutt's (2011) approach, Milne, Laney, et al. (2015) present a case for *spectral pitch class similarity* (hereafter *SPCS*) as a predictor of tonal fit. The SPCS between two stimuli is a measure of the degree to which frequencies in the spectra of the stimuli overlap (given the octave equivalence implied by the use of the scales, and uncertainties of pitch perception).<sup>4</sup> To calculate the SPCS between two stimuli, a spectral pitch class vector is first defined for each stimulus, with values representing the expected number of partials (overtones) perceived at each of 1200 log-frequency (modulo the octave) elements. Each spectral component is smeared by a smoothing width chosen upon model comparison using the data of Experiment 1, to account for uncertainties in pitch perception. Then, the SPCS between the two stimuli is the cosine similarity of these vectors (the resulting similarity lies between 0 and 1). For more detail on the calculation of SPCS, see Appendix A.

Confirmed by cross-validation, Milne, Laney, et al. (2015) show that SPCS provides a better fit to the results from Krumhansl and Kessler (1982) than a wide selection of other models. In this study, the perceived fit of any tone or chord is a function of the tones (e.g., scale) it is contextualized by. If the cognition of tonality is due at least in part to SPCS, as this suggests, then given any specific scale, differing tonal fits for tones and chords should emerge independent of prior long-term learning and without any additional cues, such as privileging certain tones and chords (or sequences of tones and chords) – by loudness, prevalence, duration, metrical

<sup>4</sup>SPCS should not be confused with *harmonicity*, which concerns instead the degree to which the frequencies in the spectra of a stimulus correspond with those of a harmonic complex tone.

weight, primacy, or recency, etc. – over others. Under this view, a context consisting of the notes (played equally often and in a random order) of a scale for which differing fits do emerge without additional cues should be functionally equivalent to Krumhansl’s carefully composed context stimuli for the major and minor scales.

Crucially, this suggests that a composer or musician sensitive to these spectral relationships may attempt to use them for musical purposes; for example, privileging high SPCS relationships and also using varying levels of SPCS to induce changing perceptions of musical fit and stability (Milne, Laney, et al., 2015). In so doing, listeners’ and future composers’ perceptions of high SPCS relationships as stable or fitting will be further enhanced by statistical learning of their higher prevalence and the typical roles they play in music. This also implies that the effects of SPCS and familiarity will be hard to distinguish because – through the above-described process – the latter captures the effects of the former along with other influences, such as artistic fashion.

After Krumhansl’s initial experiments, the probe-tone paradigm saw much application in attempts to answer different but related questions of music cognition. Through varying aspects of the design, for example the presence/absence of *tonal cues* in the context, the spectra of sounds used, the precise questions asked, and the musical sophistication of participants, the results of probe-tone experiments can take on different meanings. The effects of such differences in design are discussed in the following Sections 1.2.3 to 1.2.3, these being crucial to the precise design of our experiments as detailed in Section 2.2.

### **The context: presence or absence of tonal cues**

Through a lifetime of listening, a representation of tonal-harmonic music is said to be built in the minds of listeners. This representation may be “activated” by tonal cues. The familiar cadential chord progressions used in Krumhansl and Kessler (1982) are strong tonal cues because they are typically used in music to assert a given tonal centre. Tonal centres can also be activated with different types of cues; for example, memorability of the tonic can be enhanced through repetition and placement in temporally more salient positions (the beginning and end of a sequence, or at metrical down beats) (Deutsch, 1975, 1980), in turn likely enhancing stability. It might be true however that the observed effects of these tonal cues are simply effects of psychoacoustic phenomena and short-term memory, or, more likely, a combination

of the two such as suggested in the previous section. For probe tone experiments, the presence of tonal cues in the context does not allow these alternative interpretations of the results to be separated. When tonal cues are absent, although we can still not be absolutely certain there are no effects of statistical learning on the results, we can be confident that the possibility for such effects has been greatly reduced.

Many later probe tone experiments made no attempt to eliminate tonal cues. Some included as context stimuli either excerpts of music from an appropriate corpus (Castellano et al., 1984; Cuddy, 1993; Cuddy & Badertscher, 1987; Cuddy & Smith, 2000; Kessler et al., 1984; Krumhansl et al., 1987; Krumhansl & Schmuckler, 1986; Schmuckler, 1989; Smith & Cuddy, 2003; Toiviainen & Krumhansl, 2003), finding that goodness-of-fit ratings aligned with either the 12-TET Western tonal hierarchy, as documented by Krumhansl and Kessler (1982), or with the statistical prevalence of the probes in the corpus, if it is a Western tonal-harmonic corpus; others used a predefined grammar in a novel tuning (Loui et al., 2006; Loui et al., 2010), finding that after exposure to enough melodies using the grammar, ratings of previously unheard melodies reflected the statistics of the grammar.

West and Fryer (1990); Lantz (2002); Smith and Schmuckler (2004); and Lantz et al. (2014) conducted probe tone experiments in which the order of pitches in a diatonic (for West and Fryer) or chromatic (for the others) scale context was randomized. In the West and Fryer study the major tonic was not rated significantly higher than the mediant, dominant or subdominant of the major mode, which suggests that listeners, musically trained or otherwise, do not differentiate the major tonic as a uniquely most stable pitch class of the diatonic scale, and that ‘the time-order of notes is important to the perception of tonal hierarchy’ (West & Fryer, 1990, p. 1). Smith and Schmuckler (2004) tested the effect of total and relative duration and frequency of occurrence of pitches. They found that when the total duration of pitches in the context stimulus was distributed in proportion to the results of Krumhansl and Kessler (1982) (their *tonal hierarchy*), with the frequency of occurrence held constant across tones the perceived goodness-of-fit of the tones resembled this distribution, Participant’s ratings did not resemble a random distribution of total duration. When total duration was controlled for, frequency of occurrence distributions were not reflected in participant ratings. This does not conflict with the results of Loui et al. (2006), Loui et al. (2010), Oram and Cuddy (1995), in which frequency of occurrence of pitch was found to influence the perception of tonality, as in these studies total duration increased with frequency of occurrence.



Lantz (2002) sought to differentiate between total and relative duration, finding that relative duration was the strongest predictor of ratings, but that total duration added a significant amount of predictability. Lantz et al. (2014) varied the duration of 12 randomly ordered microtonal tones in accordance not with a Western tonal hierarchy, but with the total duration of each tone in a piece of Korean music from Nam (1998), or in a quasi-random manner. They found that when the distribution of tones in the context matched the distribution of tones in Korean music, perceived goodness-of-fit reflected the distribution of the duration of pitches in the context, for listeners both familiar and unfamiliar with Korean music. When it didn't match, 'neither group appeared to perceive a clear pitch structure beyond a long tone versus short tone distinction' (Lantz et al., 2014, p. 596). These results support the influence of relative duration of notes in the perception of tonal hierarchy, though not conflicting with aforementioned studies in which total duration, as an after effect of frequency-of-occurrence was seen to be a strong influence.

### **The dependent variable: participant ratings**

One potential issue with the variety of probe tone experiments that have been previously undertaken is the inconsistency of precisely what the participants are asked to rate, and what it is assumed that these ratings describe. Krumhansl's ratings of, initially, how well a probe "completes" a context sequence and then of how well the probe "fits" the context stimulus, have been interpreted as descriptions of tonal/-musical stability (Krumhansl, 1990, 2001). West and Fryer erroneously write that Krumhansl's participants are asked to rate tonal stability, and ask their own participants to rate their confidence in the probe tone's "suitability as a tonic". This is clearly not a comparable task to the earlier probe tone experiments (since, for the trained musician participants, this is a theoretically informed task).

In his 1997 paper, Bigand directly tests "stability", finding a larger contribution of harmony upon the perception of stability in tonal-harmonic music than of beat, rhythm, melody and phrase. Rather than using the typical probe tone paradigm in which a tone would follow a context stimulus, the stimuli were again presented in ordered fragments, starting at the beginning and concluding at each pitch in turn, from the second. The participants were asked to rate the musical stability of the final pitch of each fragment. In this way the perceived stability along the sequence

is collected. Participants were informed that ‘strong stability at the end of a fragment evokes the feeling that the melody could naturally stop at this point. On the other hand, low stability evokes the feeling that there must be a continuation of the melody’ (Bigand, 1997, p. 812). This is not too dissimilar to West and Fryer’s participants rating the suitability of the probe tone as a tonic, described as ‘a pivotal note that could round off a tune properly’ (West & Fryer, 1990, p. 255).

### **Participants: musicians or non-musicians**

The effect of musical training upon participants’ responses is an important factor to consider. Some studies have found musicians and non-musicians to give very similar responses. For example Corrigan and Trainor (2009) show that ‘Even adults with no formal music lessons have implicit musical knowledge acquired through exposure to the music of their culture’ (Corrigan & Trainor, 2009, p. 164) where ‘two of these abilities are knowledge of key membership (which notes belong in a key) and harmony (chord progressions)’ (Corrigan & Trainor, 2009, p. 164). West and Fryer found that ‘nonmusicians showed the same profile of responses as musicians’ (West & Fryer, 1990, p. 253). However, in Krumhansl and Shepard’s 1979 experiments, in which no non-musician participants were used, participants with lower levels of music experience responded more strongly to pitch height cues and gave higher ratings to the probes most similar in pitch height to the final pitch(es) of the context stimulus. This could be described as a recency effect, which has been demonstrated to significantly and positively affect memory recall in a probe tone recognition paradigm, along with primacy (Mondor & Morin, 2004), wherein the earliest item in a serial presentation is privileged in memory recall.

The probe tone experimental paradigm was also extended to chords. The following subsection reviews the literature of probe chord studies in light of the above considerations.

### **1.2.4 Probe Chord Experiments**

In addition to the probe tone experiment detailed in Krumhansl and Kessler (1982), Krumhansl (2001) includes a study of the perceived fit of probe chords to a tonal context. The context comprised an ascending and descending C or F $\sharp$  major or melodic minor scale, or a complete diatonic circle of fifths progression of triads in C major, C minor, F $\sharp$  major, or F $\sharp$  minor. This was followed by probe triads – major, minor,

and diminished – built on each note of the chromatic scale. All stimuli were presented using Shepard tones (rendering inversion meaningless). They found that ‘In a major-key context listeners strongly preferred major chords over minor and diminished chords, which were given approximately the same ratings’ (Krumhansl, 2001, p. 172). In a minor-key context the difference between the ratings of major and minor triads was smaller, and diminished triads received the lowest ratings. Krumhansl notes that this ordering reflects that of the perceived consonance of the triads. Controlling for chord type, diatonic chords obtained significantly higher ratings than non-diatonic chords. The prevalence in relevant corpora of the constituent tones of the triads, from Youngblood (1958) and Knopoff and Hutchinson (1983) provides a reasonably accurate model of the data. The tonal hierarchy of these constituent tones correlates more strongly with the data, and a model of both of these measures as well as ‘membership in diatonic set’ performs better again. When normalized for chord type, the combined model correlates well with the data: .88 for the major context, and .82 for minor.

Krumhansl also considers just the diatonic tertian triads in a second analysis. For the major context the tonic chord ‘I’ received the highest ratings, followed by ‘IV’ and ‘V’, then by ‘ii’ and ‘vi’, and finally by ‘iii’ and ‘vii°’, the triads with the weakest harmonic functions. For the minor context, the tonic ‘i’ was again the highest rated, followed by the remaining major and minor triads – ‘III’, ‘iv’, ‘V’ and ‘VI’ – and finally by ‘ii°’ and ‘vii°’. Results seem to reflect both music-theoretic predictions and differences in perceptions of chord types. For this data set a model of chord prevalence is available, using Budge’s Tables IX and X (Budge, 1943) of the frequency of occurrence of the chords in representative compositions of the eighteenth century and the first 75 years of the nineteenth century. This, along with tonal hierarchies and prevalence of the constituent tones, correlated reasonably well with the data. Krumhansl suggests that the correlation of her results with chord prevalence ‘suggests the relative frequencies of chords in the listeners’ musical experience may have initially shaped the form of the perceived harmonic hierarchy’ (Krumhansl, 2001, p. 181).

Bigand et al. (1996) required musician and non-musician participants to evaluate the ‘tension’ of a chord – a major or minor triad or a major-minor (RPCs 0-4-7-10 above the tonic, often referred to as a *dominant* seventh) or minor seventh (RPCs 0-3-7-10 above the tonic) chord built on any pitch of the chromatic scale – between two C major chords after a tonal context in C major. Using linear mixed effects

regression they find ratings of tension to depend upon a combination of tonal hierarchy (from the results of Krumhansl and Kessler (1982) or from Lerdahl's (Lerdahl, 1988) tonal pitch space distance), horizontal motion (which involves melodic continuity and recency, as we operationalize them in Sections 2.2.3 and 3.2.2) and sensory chordal consonance (involving pitch commonality, which is related to SPCS, as well as roughness), whose relative importance varies with musical training. Musical training was included only as binary variable – musicians and non-musicians – and though ordinal ratings data were collected, ordinal regression was not used.

(Bigand & Parncutt, 1999) then ran a similar experiment with a much longer context that modulated through several keys in order to test the perception of tension in long chord sequences, finding that perceived tension was due largely to local harmonic cadences and was only slightly influenced by global harmonic structure. Rather than using the typical probe tone (or chord) paradigm in which a tone (or chord) would follow a context stimulus, the stimuli were again presented in ordered fragments, starting at the beginning and concluding at each chord in turn, from the second. The participants were asked to rate the musical tension of the final chord of each fragment. In this way the perceived tension along the sequence is collected. Participants were told that 'strong musical tension at the end of a fragment evokes the feeling that there must be a continuation of the sequence. Low musical tension evokes the feeling that the sequence could naturally stop at this point' (Bigand & Parncutt, 1999, p. 242). Participants were not given this sort of information in Bigand et al. (1996). As in Bigand et al. (1996), linear mixed effects modelling was employed, with musical training – referred to this time as 'musical expertise' – included as a binary variable, i.e., musician or non-musician.

More recently, Craton et al. (2016) asked a large sample of non-musicians or amateur musicians to rate how surprising a probe chord of a major or minor triad or a dominant seventh chord was after either a tonal context or white noise, or how much they liked it. After white noise, i.e. without a tonal context, their hypothesis that ratings should reflect a rank ordering based on the relative dissonance of chord types is supported. After a tonal context they found that their results resemble those of Krumhansl (2001), as well as the frequency of occurrence of the probe chords in rock music. Whereas Krumhansl employed the use of Shepard tones in order to minimize the effects of pitch height, a piano timbre was employed in Craton's study, and no efforts to minimize the effect of pitch height were made. We should not be surprised, in that case, that they found a strong influence of pitch height on

ratings of surprise and liking.

Following this, Craton et al. (2019) ran a pair of experiments in which participants were asked to rate how much they liked major triads sounded on all 12 pitch classes after the context of an ascending and descending major scale and two C major chords. A piano timbre was used again but in order to test for the effect of pitch height the probe chords were played in two different registers. So that results could be compared more directly to other probe chord studies, while the first experiment replicated the earlier study, with participants providing liking ratings, the second experiment had participants give goodness-of-fit ratings. Ratings for both experiments were on a 10-point Likert scale; the experiments were otherwise identical. Results were similar for the two experiments, however though 'I' was given higher fit ratings than 'IV' and 'V', the liking ratings across the three chords did not differ significantly. Chords in the lower register were rated higher in both experiments. The authors' hypothesis that the basic diatonic chords (I, IV, V) would be rated above the rock-typical chords (II,  $\flat$ II,  $\flat$ III,  $\flat$ VI, VI,  $\flat$ II), which in turn would be rated above the atypical chords ( $\flat$ II,  $\sharp$ IV, VII) was confirmed. The stimulus was also played through Leman's (2000) auditory short-term memory (ASTM) model of tonal contextuality and the results were found to agree well with the fit ratings of Experiment 2 (with a Kendal's  $W$  value of 0.899, with  $p = 0.18$ ). Since the ASTM model models purely sensory processes and is dependent only upon the auditory signal, Craton et al. 'suggest the possibility that bottom-up processes create a perceptual ranking of chord fitness for all chords. This hierarchy then provides the harmonic palette from which composers/improvisers in different musical systems may conservatively (common-practice) or liberally (rock) choose' (Craton et al., 2019, pp. 16-17).

Smit et al. (2019) tested perceived pleasantness/unpleasantness and happiness/sadness of all triads from the Bohlen-Pierce system, where-in a *tritave* – an interval with frequency ratio 3/1, in contrast to the octave of 2/1 – is split into 13 equal (Experiment 2) or approximately equal (Experiment 1) steps. Smit's context stimulus was played (using a piano timbre) at a much faster rate than typical for a melody. For both experiments, Smit modelled both response types using a combination of roughness, harmonicity, spectral entropy, average pitch height and *12-TET dissimilarity*: the smallest distance to any chord of 12-TET, with each effect moderated by musical sophistication, collected via the Goldsmith Musical Sophistication Index (MSI) (Müllensiefen et al., 2014). Spectral entropy can be thought of

as an aggregation of the SPCS between all pairs of pitch classes in a pitch class set (Milne et al., 2017) – a triad, in this experiment. All predictors were found to have consistent influence in the expected direction for both response types across both Bohlen-Pierce tunings, highlighting the importance of both intrinsic and extrinsic factors on the perception of affect in music.

Though not considered to be probe triad experiments, other studies involved participants directly evaluating some quality of a small number of target chords. Like in Krumhansl and Shepard (1979), in Bigand and Pineau (1997) and Tillmann and Lebrun-Guillaud (2006) participants rated how well a target chord “completes” a musical sequence, or, in Tillmann and Lebrun-Guillaud (2006), “belongs to”. In (Steinbeis et al., 2006), participants rated the “emotionality” or “tension” of an altered or unaltered target chord within a Bach chorale, and (Corrigall & Trainor, 2009) show that children as young as three can discriminate chromatic from scale-tone chords by rating target chords as “good” or “bad”. The results of these studies are interpreted to suggest that listeners are able to discriminate chords based upon familiarity with the use of such chords in a musical corpus, but there is no reason why a bottom-up explanation for this ability to discriminate should not be possible.

Many other studies implicitly test for the perception of two different chords (not necessarily triads) – a target and a foil – after a context setting stimulus in a harmonic priming paradigm which is similar though not equivalent to a probe triad paradigm. Theoretically the context setting stimulus primes the participant to react in one way to the target triad which is congruent to the context, and in another way to the foil, which is incongruent.

### **Chord type and inversion**

In contrast to probe tone experiments where such effects are absent, chord type and inversion are important in probe chord experiments to consider in the prediction of ratings. Differing chord types and inversions thereof, lead to different chords whose consonance/dissonance/stability/tension has been shown to affect ratings of various judgements of probe chords.

As mentioned above, Krumhansl (2001) tested musically trained participants with major, minor and diminished triads, finding that triad type was significant in a model of ratings of goodness-of-fit. Bigand et al. (1996) tested major and minor

triads, and dominant and minor sevenths, finding that whilst the chord's consonance/dissonance, as modelled by roughness was significant in a model of tension ratings for musicians, it was insignificant for non-musicians. Bigand and Parncutt (1999) test for the perceived tension (described as equivalent to the inverse of stability) of chords of many different types and inversions. The tension of the chord irrespective of context, referred to as "local stability" concerned the inversion of the chord, and the inclusion of non-triad tones, but not the triad type. Local stability was found to be significant in a model of perceived tension for musician participants but not for non-musician participants. Finally, in Craton et al. (2016) a significant effect of chord type (major, minor or dominant seventh) is found for ratings both of surprise and likability. Major chords were the least surprising and most likeable, and dominant sevenths the most surprising and least likeable.

With the exception of Bigand and Parncutt (1999), in all these studies chords were played in root position<sup>5</sup>. Empirical data shows that inversion does affect consonance (Cook et al., 2007; Eberlein, 1994; Johnson-Laird et al., 2012; Roberts, 1986), but that musicians judge chords' similarity independent of the inversion (Mathews et al., 1988; Roberts & Shaw, 1984). Mathews et al. (1988) showed however that non-musicians do not judge chords of the same identity but different inversion as similar, relying only on pitch height for similarity judgements. They also found that in a novel tuning system, musicians rate the similarity of chords akin to non-musicians – according only to pitch height (It should be noted however that the tuning system employed in their study – *Bohlen-Pierce* – repeats at the interval of a *tritave* – an octave plus a perfect fifth – and therefore in this tuning inversions take on a different meaning and are not comparable to inversions in ETs and other tuning systems that repeat at an octave). This suggested to the authors that 'the ability to abstract more complex information depends on training' (Mathews et al., 1988, p. 1214).

### **Tension and stability**

Given that Bigand and Parncutt (1999) suggested that tension can be considered equivalent to the inverse of stability along with Krumhansl's (1990, 2001) assumption that goodness-of-fit provides a measure of musical stability we can expect ratings of stability to be similarly affected by triad type and inversion. The possible

---

<sup>5</sup>Krumhansl's use of Shepard tones for her stimulus renders inversion meaningless, so it is not entirely correct to say that they are in root position, however like the other studies inversion is not included as a predictor in a model of probe chord ratings.

equivalence of stability and consonance has also been suggested by Cook and Fujisawa (2006) and is described by Parncutt and Hair (2011). Further, Eberlein (1994) found that the sequence of perceived consonance of triad types (major > minor > diminished > augmented) was reflected by their prevalence in a tonal-harmonic musical corpus. The prevalence of the notes of the triads within a tonal-harmonic corpus was used by Krumhansl, however, in addition to chord type as a predictor of perceived goodness-of-fit. Cook (2009) considers stability to be due to a combination of consonance and tension/sonority. He considers tension in diminished and augmented triads that is absent in major and minor triads to be due to the intervallic equidistance of the three notes of the triad (assuming root position for diminished triads), a “three-tone effect” whereas consonance/dissonance is “two-tone effect” due to the sensory dissonance of the intervals making up the triads. Bigand et al. (1996), however, showed that tension involves a combination of sensory consonance/dissonance, tonal hierarchy inside a musical context, and horizontal motion, where tonal hierarchy is modelled by the goodness-of-fit rating of the pitch-classes the chord contains after a tonal context from Krumhansl and Kessler (1982), or by Lerdahl’s tonal pitch space distance (Lerdahl, 1988). In Bigand and Parncutt (1999) tonal function is used in models instead of tonal hierarchy, and instead of consonance/dissonance a combination of local stability and pitch commonality – related to SPCS, but simpler – is used.

What is described as tension in the above studies seems to match Johnson-Laird’s conceptualisation of dissonance. After a survey of models of consonance/dissonance, he introduces a “dual-process model” in which dissonance consists of both sensory dissonance (largely comprising roughness) and “tonal dissonance”, described as ‘the high-level cognitive processes that rely on a tacit knowledge of the principles of tonality’ (Johnson-Laird et al., 2012, p. 23). Two experiments were run testing the perceived consonance, which is described as “pleasantness” in instructions for the benefit of the non-musician participants, and equated with both pleasantness and stability in the paper’s introduction. A 7-point Likert scale was used to collect ratings of triads (Experiment 1) and tetrads (Experiment 2). The dual-process model was found to be a significantly stronger predictor of ratings than sensory dissonance alone. A third experiment confirmed the hypothesis that consonant chords are rated as more consonant when they occur in a tonal sequence, whereas dissonant chords are not reliably affected by this manipulation.



Though terms like stability and tension are not defined consistently across the literature, we can be sure that ratings of some quality of chords played after a context-setting stimulus involve effects intrinsic to the chord itself, which may include consonance/dissonance, roughness, local stability, triad type, inversion and intervallic equidistance, as well as context dependent effects which may include pitch commonality, SPCS, prevalence and horizontal motion. These effects may or may not vary for a particular experimental design.

### 1.2.5 The cognition of music in novel/microtonal tuning systems

Research into the cognition of music in novel tuning systems has been relatively sparse. In addition to Loui and Wessel (2008) and Leung and Dean (2018), Loui (2012) found that while removing small steps from the melody greatly inhibited learning, the consonance of the scale affected preferences but had no effect on learning.

Parncutt and Cohen 1995 also found an effect of step size on the recognition of microtonal melodies. They found that, from 100c (an equal tempered semitone; 1/12th of an octave), when step sizes were decreased to 40c, no change occurred, but from 30c onwards, recognition performance decreased. This research suggests a lower limit of 30c for step sizes in microtonal scales used for tonal-harmonic music production. Strasburger and Parncutt (1994) found that mistuning a note of a chord up to 50c strongly affects the *virtual pitch* (the perceived pitch of a *complex pitch* – not a sine tone – corresponding to the lowest frequency – fundamental – in a harmonic series suggested by the spectral components of the pitch). This suggests an upper limit of 24 for the possible number of steps per octave in a tuning system. Alternatively, ‘being safe’ by allowing up to 50c mistuning between notes leads to approximately 12-TET as a limit. This suggestion supports Krumhansl’s 1979 results wherein quarter tone pitches in between the pitches of 12-TET from which the context stimuli was taken, when probed were not perceived as harmonically distinct; they were given a fit rating averaging the ratings of their closest 12-TET neighbours. Bailes et al. (2015) similarly found that non-musicians were unable to categorically distinguish quarter tone intervals from neighbouring 12-TET intervals, but that musicians were able to make this distinction.

Strasburger and Parncutt (1994) suggested that the width of perceptual pitch

categories is understood not to be immutable, though learning of microtonal systems would only be effective early in life. Zatorre et al. (2012) showed however that training can induce rapid improvement in adults. Before training all participants were able to discriminate between melodies consisting of intervals of 20c, and after, of 10c. As my experiments will involve all tones of 22-TET (a step size of approximately 55c) I will be able to test the validity of some of these contrasting predictions. If participants were found to be unable to distinguish between neighbouring notes of 22-TET at any stage this will have been noted, and scales employed in future experiments would not include consecutive steps of 1 step of 22-TET, however no such problems were uncovered.

Though no experiment has tested the perception of music in 22-TET, Bucht and Huovinen (2004) use the forced choice method to empirically measure the consonance of intervals in 19-TET, a historically popular alternative tuning system (Mandelbaum, 1961). The authors reason that various strategies are used in judgment of consonance, largely sensory consonance, and less so, elimination of slow beating and (familiarity of) fundamental frequency relations. Their paradigm compares the consonances of neighbouring intervals. I will not employ a forced choice paradigm as I require direct comparisons between all notes and triads. Mathews et al. (1988) tested the perceived consonance of all possible triads of the Bohlen-Pierce scale, the same tuning system later employed in Loui and Wessel (2008), Loui (2012), and Smit et al. (2019). Mathews et al. (1988) observed a wide range of consonance and found triads which included a 1-step interval to be most dissonant. They suggest that roughness models fit the data well, but did not test harmonicity models, as modern harmonicity models had yet to be developed. They also asked participants to rate the similarity of chords and their inversions. They found that for musicians' ratings for chords of 12-TET were influenced by key relationships, inversions and chord type. Ratings of Bohlen-Pierce chords were dependent however only upon pitch height, where chords including notes at a similar pitch height were judged to be similar. Non-musicians judged both traditional and non-traditional chords only by pitch height. This highlights the influence of learning upon music cognition.

Milne and Holland (2016) conducted an experiment in which participants were presented with a series of melodies in a variety of tuning systems, each available in two varieties: The spectra of tones used in the melodies either matched or did not match the tuning used for the melody. They were asked to choose which melody exhibited the greatest overall affinity between tones. Milne's results supported his

hypothesis that affinity is impacted both by the spectral pitch similarity (psychoacoustic pitch similarity) between successive tones and the harmonicity of each tone. Milne suggests that his affinity model, relying upon only bottom-up processes, is able to explain historic and contemporary scale structures commonly found in music because they also exhibit relatively high overall SPCS (Milne & Holland, 2016).

As stated in the introduction, my research will be the first empirical study of the cognition of tonality in microtonal scales. By using a novel tuning system effects of familiarity will be weakened and I will be able to provide a stronger investigation of bottom-up models for the cognition of harmonic tonality. For a scale to be able to support harmonic tonality the chords it contains, as well as the RPCs, must fall into a hierarchy of perceived stability. The presence of such a hierarchy is tested for in the RPCs and tertian triads of each scale considered. In the case that a unique most stable RPC or triad exists, it can be considered the tonic of the scale, though we do not suppose that a scale cannot support harmonic tonality without the existence of a unique tonic. Our experiments also comment on the use of triads in the establishment of tonal hierarchy.

## Chapter 2

# Experiments 1 & 2 – Tones

### 2.1 Introduction

The first two experiments tested the perceived fit and stability of tones and triads of 9 scales. The first experiment comprised three commonly used scales as context: Diatonic, harmonic minor and jazz minor (melodic minor ascending). One block tested all 12 pitch classes of 12-TET as probes and a second block tested all tertian triads of the scales, along with a selection of additional triads, after a randomly ordered, isochronous sounding of the notes of the context scales. In an effort to reduce the possible effect of familiarity, the second experiment comprised 6 other scales as context, for which no Prevalence data are available, namely, the harmonic major, double harmonic, pentatonic, hexatonic, octatonic and blues scales. Other than pentatonic and perhaps the blues, these scales are far less commonly used than the three scales of experiment one. In this experiment a probe triad block tested the perceived fits and stabilities of all 12 pitch classes of 12-TET after a randomly ordered, isochronous sounding of the notes of all six context scales. Since tertian triads are defined only for 7-note scales, the probe triad block comprises only the 7-note scales, namely, the harmonic major and double harmonic. Of these scales, the tertian triads that are either major, minor, diminished or augmented are probed after a randomly ordered, isochronous sounding of the notes of these two scales.

This chapter and the next detail these two experiments. This chapter introduces the experimental procedure and analysis techniques that we use throughout all experiments in this thesis, and details the results of the probe tone block, and the following chapter details the results of the probe triad blocks.

## 2.2 Overview of the Experiments and Models

### 2.2.1 Experimental design

The first experiment was registered as part of the Open Science Framework's pre-registration challenge, and accordingly the method, hypotheses and intended analysis at the time of preregistration, which we have attempted to maintain as closely as possible, can be found online at <https://osf.io/az6x8>. Two probe tone experiments were run in order to test for the perceived stability and goodness-of-fit of relative pitch classes given the context of a number of different scales. In order to examine effects of scale structure on the perceived stability of scale pitches, the duration and frequency of occurrence of each scale pitch is kept constant. Similarly to West and Fryer (1990); Lantz (2002); Smith and Schmuckler (2004); and Lantz et al. (2014), to explicitly avoid tonal cues in our experiment, the order of scale pitches and the relative pitch classes of the lowest (and therefore highest) pitches of the scales are randomized, along with the overall pitch height of the stimulus, with all randomizations independent of each other.

As mentioned in Section 1.2.3, in Krumhansl and Kessler (1982) and Krumhansl (2001), Shepard tones were used to disguise pitch height after some participants were found to respond largely to pitch height cues. As Shepard tones do not closely resemble the spectra of real musical instruments, and we seek to employ physically plausible and commonly occurring types of sounds, we used harmonic complex tones in our experiments. Effects of pitch height at an individual observation level are controlled for in our model and should average out at a population level, due to the randomizations described in Section 2.3.1. Whereas, for example, West and Fryer (1990) employ a piano timbre and Smith and Schmuckler (2004) use an electric piano timbre, we opt for a simpler timbre (one that can be described by a single parameter, so that it can be more easily taken into account in analysis) that is generally familiar in quality but not reminiscent of any particular instrument, such that results may be more generalizable. Our timbre, explained in more detail in Section 2.3.1, is created with additive synthesis and resembles a saw tooth wave under a low pass filter.

Our experiments also explore the appropriateness of equating fit and stability. Half the participants were asked to rate the goodness-of-fit of the probe to the context; while the other half rated the stability of the probe given the context, informed

that ‘a musical sound is considered to be stable if it does not need to move (resolve) to another musical sound’ – an explanation that should be understandable to both musicians and non-musicians.

Both musicians and non-musicians are recruited such that a broader spectrum of musical sophistication is tested, and the results may more easily apply to any hearing person. In the analyses of the resulting data, rather than dividing participants into groups – as in Krumhansl and Shepard (1979); and West and Fryer (1990) – we collect participants’ responses to the Goldsmith Musical Sophistication Index (MSI) (Müllensiefen et al., 2014) questionnaire in order to obtain a continuous measure of musical sophistication to use as a potential predictor in our model, expecting it to interact significantly with other variables.

Considering the randomization of the order of tones of the context stimulus in our experiments, after averaging over all trials for each context scale and probe combination we should expect no “non-tonal” effects (effects unrelated to a pitch’s tonal function within a scale, e.g. pitch height and recency) to influence our average ratings; we expect to see influence of such effects only on a per-trial basis.

“Pleasantness” or “pleasingness”, considered comparable to stability by some researchers (Cook & Fujisawa, 2006; Cook et al., 2007), has been shown to be inversely proportional to frequency for sine tones (Berlyne et al., 1967; Guilford, 1954; Parham, 1987) across the frequency range of our stimulus. This suggests that we should expect an effect of absolute pitch height. The pitch height effect observed in Krumhansl and Shepard (1979) may relate, as well as to recency, to pitch height relative to the average pitch height of the context melody. Pitch height, relative pitch height, primacy and recency are included as effects in our model, as well as the squares of pitch height and relative pitch height, to allow for a non-linear relationship, where mid-range probes or probes closer to the centre of the context stimulus may be rated above others.

Also included in our model are effects of trial number and block order. Considering that task performance may increase over trial number through familiarisation with the task, but may then also decrease again with fatigue, we include predictors of both trial number and trial number squared to allow for a non-linear effect. Further, we test for the effect of the frequency of occurrence of the pitch of the probe within the experiment up until the time of that probe (referred to as its *Count*), because this has been demonstrated in Loui et al. (2010) to affect goodness-of-fit ratings in probe tone experiments within the time-frame of a single-session experiment.

Furthermore, results from Loui and Wessel (2008) and Leung and Dean (2018) suggest that learning of a musical system can occur rapidly (within a reasonable time frame for an experiment). As detailed in Section 2.3.1 below, we keep trials of each scale together in blocks order to allow such learning effects if they may occur. To test for these we use as an interaction effect in our model the *within context trial number* – a trial number count that resets to 1 at the introduction of each new context scale. Our model also includes something we call *melodic continuity*, where we might expect that if a rising or lowering pitch sequence precedes the probe, higher ratings will result when the sequence continues through to the probe (regardless of the interval sizes). Finally, given that a serial response effect has been observed in similar (but not probe tone) experiments (for example, Dyson and Quinlan, 2010), we include in our model an effect of the rating given to the previous trial on the rating for the current trial. Our measure of recency – whether or not the probe pitch is equivalent to the final pitch of the context – may also capture this effect.

We ran two experiments which differed mostly by which scales were used in the context stimulus. The first included the three most common and important scales in tonal-harmonic music: diatonic, harmonic minor and jazz minor (melodic minor ascending). We expect SPCS and Prevalence to together predict our data, with effects such as pitch height also influencing ratings for individual trials. *Prevalence* is represented by the frequency of occurrence of relative pitch classes within an appropriate tonal-harmonic corpus (as explained in Section 2.2.3 below), and is used as a measure of familiarity. A second experiment includes an additional six scales, four of which are less common in Western music and for all of which Prevalence data are unavailable, namely: pentatonic, harmonic major, double harmonic, blues, octatonic and hexatonic.

Whereas in Experiment 1 all scales comprise 7 notes, Experiment 2 comprises scales of 5, 6, 7 and 8 notes. Accordingly, the number of notes in the scale is also used a predictor in the analysis for this Experiment.

## 2.2.2 Hypotheses

### Experiment 1

Given Krumhansl’s assumption that goodness-of-fit ratings directly measure musical stability, we should expect that our ratings of stability largely resemble those of goodness-of-fit. Pilot data suggested however that some particular pitches such as

the leading-tones of the scales exhibit significantly lower stability than goodness-of-fit. This observation is consistent with music-theoretic ideas, where the leading-tone of a scale leads strongly to the tonic, rendering it very unstable, but is still a member of the scale, so may fit the context reasonably well. Accordingly, our first hypothesis was that:

H1: Ratings of stability differ insignificantly from ratings of goodness-of-fit, apart from in a small number of cases that reflect music-theoretic ideas or tonal-harmonic musical practice.

We expect SPCS to model the ratings well, but we assume that enculturation will also affect results. Accordingly, our second hypothesis was:

H2: Perceived goodness-of-fit and perceived stability of probe tones<sup>1</sup> may be modelled by the SPCS of the aggregated pitches of the context and the probe pitch, and the statistical prevalence in Western music of the probe within the context scale.<sup>2</sup>

Finally, considering that Krumhansl found that participants with less musical training responded more to pitch height cues, we expect that the musical sophistication of the participants will affect the degree to which they respond to the predictors in our model; that is,

H3: Significant interaction effects exist between musical sophistication and other predictors in such a model.

## Experiment 2

Considering the scales used in Experiment 2 are less familiar overall, we do not expect to see differences between goodness-of-fit and stability ratings for specific pitch classes. Our first hypothesis is thus simplified:

H1: Ratings of stability differ insignificantly from ratings of goodness-of-fit.

Given that we cannot test for Prevalence for these scales (and, due to the unfamiliarity of scales would not expect it to effect out results as strongly as in the first experiment even if we were able to test it), our second hypothesis reads:

H2: Perceived goodness-of-fit and perceived stability of probe tones may be modelled by the SPCS of the aggregated pitches of the context and the probe pitch.

Finally, our third hypothesis remains from Experiment 1:

<sup>1</sup>In the preregistration our hypothesis mentioned probe triads as well as tones. Discussion of the probe triads tested in this experiment can be found in Chapter 3.

<sup>2</sup>In our preregistration we also included the consonance/dissonance of the probe as a hypothesized predictor. This only varies in the probe triad block, which is discussed Chapter 3.



H3: Significant interaction effects exist between musical sophistication and other predictors in such a model.

### 2.2.3 Analysis

For both experiments, we test initially for each scale using model comparison whether the RPC of the probe is a significant predictor of ratings – i.e., whether or not a tonal hierarchy is observed for average ratings (whether fit or stability). We then test Hypothesis 1, concerning the similarity of fit and stability ratings. Following this is a descriptive model of the data concerning the accuracy to which SPCS is able to predict ratings, averaged over all trials, for each probe-context combination (this descriptive model is not used to test hypotheses). Hypotheses 2 and 3 are tested using a model of by-trial predictors for ratings including things like pitch height, recency and trial number, as well as the key variables of interest – Prevalence and SPCS. For Experiment 1 this model is labelled *Model 2.1*, and for Experiment 2 it is labelled *Model 2.2*.

The experiment was preregistered with the Open Science Framework as part of the preregistration challenge, available at <https://osf.io/az6x8>. After this study was preregistered, we realised that the analysis could be potentially improved by the inclusion of a small number of effects not included in the pre-registration. We believe that considering these added effects allows for a fuller description of the data. Inferences made from the hypothesis tests on the preregistered model, as shown in Appendix B, Table B.1, are the same as for Model 2.1. In the main text the hypothesis tests are run for Experiments 1 and 2 on Models 2.1 and 2.2 respectively. As Experiment 2 was not preregistered, no associated pre-registered model is run.

Model 2.2 is re-run for the combined data sets for Experiments 1 and 2, and is labelled *Model 2.3*. Finally, exploratory analyses – using extensions of Models 2.1, 2.2, and 2.3 are included.

Section 2.2.3 below details each of the above tests and models, as well as the latter's predictors.

#### Bayesian ordinal mixed effects models

Bayesian regression is used to test Hypotheses 2 and 3, and to test for an observed tonal hierarchy. The R package *brms* (Bürkner, 2017, 2018) is used to conduct a Bayesian ordinal mixed effects regression for the ratings (either fit or stability) of all

individual trials. Ordinal (cumulative logit) regression is used because of the ordinal nature of the dependent variable (Likert ratings of either fit or stability). Mixed effects are used such that the intercept and slopes can vary between participants. These are returned as population-level (fixed) effects and participant-level (random) effects.

Rather than a point estimate of each predictor's most probable effect, Bayesian regression calculates the whole posterior probability distribution of each predictor's effect, given the observed data and a prior distribution. This allows for *credibility intervals* to be calculated. The 95% credibility interval of an effect is the interval that we can be 95% certain contains the effect's true value. Accordingly, credibility intervals have a more straightforward and intuitive meaning than confidence intervals in classical regression. Bayesian regression also enables the calculation of *evidence ratios*, the odds (probability ratios) in favour of directional hypotheses (such as a given effect being greater than zero).

For example, if the integral of the posterior distribution over the interval  $(0, \infty)$  is  $p$ , the evidence ratio in favour of the effect being greater than 0 is  $p/(1 - p)$ ; so, if the lower boundary of a (one-sided) 95% credibility interval is precisely zero, this implies that there is a 5% probability the effect is less than zero and a 95% probability it is greater than zero; hence, the evidence ratio is  $.95/.05 = 19$ . (Stanford et al., 2018, pp. 9–10)

We followed the guidelines proposed by Jeffreys (1998), cited in Dienes and Mclatchie (2018), Kruschke and Liddell (2018), in order 'to qualify the weight of evidence for or against any given hypothesis (e.g., that an effect is greater than 0)' (Stanford et al., 2018, p. 10). The guidelines proposed can be summarized as thus: Evidence ratios of 1–3 suggest that no evidence for the tested hypothesis exists; evidence ratios of 3–10 suggest "moderate" evidence for the hypothesis; evidence ratios of 10–30 suggest "strong" evidence; and evidence ratios above 30 suggest "very strong" evidence (Stanford et al., 2018).

Models are compared through the *Pareto smoothed importance sampling (PSIS)* approximation for the *leave-one-out cross validation information criterion*, or *LOOIC*, which is a measure of how well a model predicts out-of-sample data, for which lower values indicate better performance. A difference in LOOIC is considered to be significant if it is at least 2 times the estimated standard error (SE) involved in

the comparison, as long as the LOOIC values are over 8, and there are at least 100 samples Vehtari, 2020.

All continuous independent variables were *standardized* – centered at 0 and scaled to have a standard deviation of 1. This reduces multicollinearity in squared terms and helps to ensure conditional main effects are interpretable (Gelman & Hill, 2006). The model was run in *brms* in *R* using

$$\text{student\_t}(3, 0, 2.5)$$

(a *t*-distribution with 3 degrees of freedom, with mean of 0, scaled by 2.5) as a weakly informative prior (a weakly informative prior reflects the fact that the researcher does not have a strong basis for prior expectations and considers the null hypothesis of zero effect size to be the most probable). The suitability of this prior is enhanced by the choice to standardize the predictors.

To interpret the resulting effect sizes: Though the data are ordinal, we assume it overlies a latent continuous variable. As we are using a logit link function, the latent variable follows a logistic distribution (Agresti, 2010) and the units of the latent variable are standard deviations of this distribution. When an observation crosses a threshold or cutpoint in the latent variable it moves up a step in the ordinal value. The models' intercepts indicate the cutpoints or thresholds in the continuous latent variable. Effect sizes in the model represent the units by which the latent variable changes for an increase of one standard deviation in the value of the predictor.

### **Test for tonal hierarchies**

Before the hypotheses are tested however, to test for the emergence of a tonal hierarchy from ratings for a scale we first run and compare two simple Bayesian ordinal mixed effects models. The models differ from those above in their predictors of ratings, which, for the first model consist only of the RPC of the probe, and for the second model only of the intercept (in both cases as both fixed and random effects). The two models are compared via cross-validation. If the model with probe significantly outperforms the model of the intercept (the null model) then this suggests that a tonal hierarchy was observed for the scale. There must be at least two hierarchical levels within the scale-tones for there to be a tonal hierarchy.

### Comparison of fit and stability ratings (H1)

A Mann-Whitney  $U$ -test was run to compare the fit to the stability ratings for each of the 36 scale-probe combinations for Experiment 1, and 55 scale-probe combinations for Experiment 2. For Experiment 1 we hypothesized that fit and stability ratings differ significantly only for a music-theoretically appropriate selection of RPCs. The hypothesis is confirmed if, after Bonferroni corrections, RPCs with significantly ( $p < .05$ ) different fit and stability ratings do not outnumber those without, and align simply with music-theoretic discourse (e.g., they are leading tones in the scale). Plots of bootstrapped fit and stability ratings for each probe-tone combination accompany the  $U$ -test results. The corresponding hypothesis for Experiment 2 is supported if no significant differences are found. As mentioned above, H1 concerns the individual scale-probe combinations for comparisons of fit to stability rather than overall differences in the predictors' effects between these two types of ratings, which are assessed using comparisons of alternative versions of the Bayesian mixed effects model.

### Descriptive model

Before Hypotheses 2 and 3 are tested, SPCS is used to predict the average ratings for each scale-probe combination under the assumption that fit and stability do not differ significantly. In this way SPCS may be more directly compared to existing models of probe tone ratings and it makes possible some useful summary visualisations. The SPCS of two pitch class sets is the cosine similarity between their spectral components (given the octave equivalence implied by the use of the scales) after the application of Gaussian smoothing (which accounts for inaccuracies of human pitch perception). A smoothing width of 10 cents (10% of a 12-TET semitone) was used. Smoothing widths of 6 and 14 cents were also tested but resulted in a marginally less well-fitting model; furthermore, 10 cents is close to values previously optimized in related experiments such as Milne, Laney, et al. (2015) and Milne et al. (2016), and was used in an SPCS model to predict perceived change in sound- as well as note-based music in Dean et al. (2019). Appendix A details the calculation of SPCS. For information further to this see Milne, Laney, et al. (2015, 2016), Milne et al. (2011); and Milne and Holland (2016).

### Testing H2 & 3 with Bayesian ordinal mixed effects models

Considering our first experiment as preregistered: following the recommendation of Barr et al. (2013) we ran the models with the maximal random effects structure driven by the design of the experiment, including random effects on participants with respect to SPCS, Prevalence, Primacy, Recency, Melodic Continuity, Count, and Relative Height, and their correlations.

Population-level effects considered in the preregistered model for Experiment 1 are:

- *SPCS*: The spectral pitch class similarity of the aggregated pitches in the context and the probe.
- *Prevalence*: The frequency of occurrence of the probe in a corpus appropriate for the context scale. Detailed below.
- *Recency*: Coded as 1 when the pitch of the probe matches the final pitch of the context, and 0 otherwise.
- *Primacy*: Coded as 1 when the pitch of the probe matches the initial pitch of the context, and 0 otherwise.
- *MelCont*: The melodic continuity of the probe from the context – the number of consecutive intervals in a single direction that can be traced back from the probe (may take integer values from 1 to 7, given the one octave range).
- *Previous*: The rating given to the previous trial.
- *Count*: The number of occurrences in the experimental stimulus of the pitch of the probe up until the point at which the probe is heard.
- *RelHeight*: The pitch height of the probe relative to its context (measured in semitones above the highest pitch of the context. May take integer values from -12 to 0 as the probe can be a semitone lower than the lowest pitch of the context).
- *RelHeight<sup>2</sup>*: The relative pitch height of the probe squared.
- *Height*: The pitch height of the probe.
- *Height<sup>2</sup>*: The pitch height of the probe squared.
- *TrialNo*: Trial number.
- *TrialNo<sup>2</sup>*: Trial number squared.
- *InContTrialNo*: Trial number within the group of trials of the same context scale.

- *Task*: Coded as  $-0.5$  if the participants rate fit and  $0.5$  if the participants rate stability.
- *BlockOrder*: Coded as  $-0.5$  if the probe tone block is presented first and  $0.5$  if the probe triad block is presented first (the data concerning the probe triad block are not considered here).
- *MusSoph*: The musical sophistication of the participant, as measured by the Goldsmith Musical Sophistication Index.

Height<sup>2</sup> and TrialNo<sup>2</sup> are included in Model 2.1 in order to enable the modelling of a single bend in the distribution of the effect of Height and TrialNo (or of interactions with those effects) on ratings.

A model comparison (via LOOIC) revealed that the removal of the effect of Task from the preregistered model does not reduce the performance of the model. Accordingly, it is not included in Model 2.1, which without this effect then can predict ratings of both fit and stability, and is considerably simpler.

Model 2.2 differs from Model 2.1 in two ways: Model 2.2 does not include Prevalence, as values for such a variable for the scales of Experiment 2 are unavailable. Given that the stimulus for Experiment 2 comprises context scales of 5, 6 and 8 notes, as well as 7, it also includes

- *ScaleSize*: The number of notes in the context scale

In the preregistration (for Experiment 1. More details follow under Section 2.3 below), we intended to test all possible two-way interactions between all these variables. Such a model would be unfeasibly complex, both to computationally fit and to understand. Accordingly, we split the variables into two groups – those which represent a feature of the stimulus, and those which may affect the relative influence of such a feature on the participant’s rating. Each variable of the second group interacts with each variable of the first group. The second group consists of MusSoph, TrialNo, TrialNo<sup>2</sup>, InContTrialNo, Task, and BlockOrder; the first group comprises the remaining effects.

Prevalence is more complex than any other effect in our model. As discussed above, the earlier probe tone experiments included ‘context setting stimuli’ assumed to activate a psychological representation of tonal-harmonic music in the listener in a particular key. It is against this representation that the participant is expected to respond to the probe. In our experiment all features that may cue such an induction are removed, apart from the RPCs of the scale. Though sections of tonal-harmonic

music frequently favour each of the scales used in our context stimuli, additional out-of-scale pitches are commonly used and these do not necessarily induce a perceived change of key. This suggests that, for example, the RPCs of the harmonic minor scale may not faithfully represent a minor-key context.

Despite this, there is a principled way in which we may proceed that should still be useful. Our experiment considers three scales: diatonic, harmonic and jazz minor. Use of the diatonic scale most frequently involves major tonality, built on the (major) tonic of the scale's major mode. The harmonic minor and jazz minor scales are more frequently associated with a minor tonic. Minor tonality is associated with the minor mode of the harmonic minor scale and of the melodic minor scale, which in its ascending form is equivalent to the exemplar mode of the jazz minor scale (provided in Section 1.1.1 in Chapter 1). As confirmed by our data (see Figures 2.4–2.6, the most stable/fitting RPC of the diatonic scale is its major tonic and the most stable/fitting RPCs for the harmonic minor and jazz minor scales are their minor tonics. Accordingly, the results for the diatonic scale will be modelled by the prevalence of the RPCs within an appropriate major-mode tonal corpus, and for the harmonic minor and jazz minor scale, by the prevalence of the RPCs within an appropriate minor-mode tonal corpus.

But this raises an additional question – what corpora are appropriate? Listening background questions included in the questionnaire completed by all participants suggest they listen to pop/rock music much more than classical music (out of 63 participants, 26 listed rock and/or pop music compared to 7 classical, 9 both and 21 other). Accordingly, it makes more sense to build our model from the prevalence of RPCs in rock/pop music corpora than from classical music corpora such as those used in (Krumhansl, 2001) and elsewhere.

De Clercq and Temperley (2011) introduced the *RS 5x20* corpus, consisting of the top 20 songs on the Rolling Stone magazine's list of the "500 Greatest Songs of All Time" from each decade from the 1950s through the 1990s (minus one, which was removed due to an absence of triadic harmony). In a later paper they added the next 101 songs from the list, forming the *RS200* corpus. The statistical prevalence of chords relative to a tonic within this corpus was found to be comparable to observed liking ratings of probe chords in Craton et al. (2016). Temperley and de Clercq experimented with natural clustering of songs from this corpus given their melodic and harmonic make up. Exploring whether or not the common-practice

major/minor tonal system applies to pop/rock music, they find that a 2-cluster solution separates songs into what seems like major, and a cluster that, though not particularly representative of minor, favours the minor third over the major third above the tonic (Temperley & de Clercq, 2013). Considering that even in common-practice Western music minor tonality is not limited to either the harmonic minor or jazz minor scales, but something more complex (e.g., the  $b2$  and  $b7$  are commonly used in the minor tonality but not found in either of those two scales), and that our highest rated RPC from the diatonic scale is the ‘major tonic’ and the highest rated from the harmonic minor and jazz minor scale the ‘minor tonic’, we use the statistical prevalence of RPCs in the ‘major’ cluster for melody to model our probe tone data from the diatonic scale, and the statistical prevalence of RPCs in the other cluster for melody to model our probe tone data from the harmonic minor and jazz minor scales. Though these are plotted in Temperley and de Clercq (2013), the values were not listed, and were provided to us via personal communication (D. Temperley, personal communication, May 12, 2018).

For each model, the population-level effects that are *significant* in the models, along with their conditional effects and the intercepts, are displayed in a table. Here, we consider an effect to be significant when its 95% credibility interval does not cross the 0 line. If the interval lies above the zero line this means that the effect is at least 95% likely to be positive and if below, 95% likely to be negative (corresponding to evidence ratios  $> 19$ ).

*brms*'s *hypothesis test* is also run for all significant effects in the models to quantify their evidence ratios. For positive effects we test the evidence ratio supporting the hypothesis that the effect lies above zero, and for negative effects that it lies below zero. These evidence ratios are listed after reported effects as “evid. ratio”. To visually test for any systematic discrepancies between the observed data and the model predictions (Gelman et al., 2013) *brms*'s *ppcheck* (a *posterior predictive check*) is used, ‘simulating replicated data under the fitted model and then comparing these to the observed data’ (Gelman & Hill, 2006, p. 158), in order to test whether or not the model makes reasonable predictions. A Bayesian version of the McKelvey-Zavoina pseudo- $R^2$  value (McKelvey & Zavoina, 1975) is also calculated, approximating the  $R^2$  value for model fit that would have been obtained if a linear model had been run on observations of the continuous latent variable underlying the discrete responses (Hagle & Mitchell, 1992; Veall & Zimmermann, 1992, 1994).

To remind the reader, we hypothesize (H2) that perceived goodness-of-fit and



perceived stability of probe tones may be modelled by the SPCS of the aggregated pitches of the context and the probe pitch, and (for Experiment 1 only) the statistical prevalence in Western music of the probe within the context scale. To answer H2, an associated reduced model is run, and the full model and the reduced model are compared via cross-validation. In Experiment 1, the reduced model differs from the full model only by the absence of SPCS and Prevalence, and for Experiment 2 only by the absence of SPCS (since the models of Experiment 2 do not include Prevalence). The hypothesis is confirmed in each case if the reduced model is significantly outperformed. H3 concerns the significance of the interaction of several effects with musical sophistication in the model. The hypothesis is confirmed if the 95% credibility interval lies entirely above or below zero for one or more interaction effects with musical sophistication.

### Exploratory analysis

In order to explore whether SPCS accounts for any differences in ratings more complex than simply whether or not the probe was included in the context, in which it could be seen as only a short term memory model, an exploratory analysis considered an additional effect of *ScaleTone*, coded as 1 if the probe tone is in the context melody, and as 0 otherwise. This consideration was explored by adjusting Models 2.1, 2.2 and 2.3 by replacing SPCS with *ScaleTone* and comparing via cross-validation to Models 2.1, 2.2 and 2.3 respectively. A model equivalent to Model 1, but without Prevalence was also compared with a model equivalent to it but with *ScaleTone* instead of SPCS.

## 2.3 Experiment 1: Diatonic, Harmonic Minor, Jazz Minor

The first experiment tested the perceived goodness-of-fit and stability of tones given the randomly ordered, uniformly distributed sounding of the pitches (sounded three times each) of three different context scales common to Western tonal music – the diatonic, harmonic minor and jazz minor – sounded with harmonic complex tones. If we find that the results cannot be modelled accurately by SPCS, this psychoacoustic description of the cognition of tonality is not supported.

### 2.3.1 Method

#### Participants

Thirty-two musicians (participants who reported having received 5 or more years of music experience) and 32 non-musicians were recruited for the experiment. One non-musician participant's data were removed after participation due to it being partially lost. The following refers to the 63 remaining participants. Non-musician participants were university students (mostly first-year) recruited through Western Sydney University School of Psychology and Social Science's SONA system and received credit points towards their degrees for their participation. Musician participants were recruited via personal connection and received a \$30 reimbursement for their time and travel to the university campus. All participants reported normal hearing capabilities. Nineteen (musician) participants reported having received 10 or more years of musical training and four reported having absolute pitch. Participants had a mean age of 27.3 years, with a SD of 10.6 years. Of the 32 musicians, 7 were female, and of the 31 non-musicians, 25 were female. This research was approved by the Western Sydney University Human Research Ethics Committee under the number H11908.

#### Stimulus

Context stimuli and probes were all sounded as harmonic complex tones with 35 harmonics (maximising the number of harmonics given the constraint that the highest harmonics of the highest pitch does not exceed the Nyquist frequency of 22,050 kHz), where the amplitude of each harmonic  $n$  is equal to  $1/n^{5/3}$  times the amplitude of the fundamental, which is fixed for all frequencies. Such a timbre resembles that produced by a saw-tooth wave, which approximates the timbre of string and brass instruments. The ratio of 5/3 is chosen for the roll-off (rather than 1, as is the case for saw-tooth waves) as it was found to be a simple representation of a roll-off low enough such that the timbre sounds distinctly different to a sine wave, but high enough that there is not too much energy in the upper harmonics and the resulting sound is not too "shrill" and unpleasant to listen to.

Context melodies consist of 3 soundings of each pitch of the 7-note context scale, in a random temporal order (i.e., 21 items taken randomly without replacement). Context scales include the diatonic, the harmonic minor and jazz minor (melodic

minor ascending) scales, tuned to 12-tone equal temperament (12-TET). The pitch classes of the scales are fixed within participants and randomized between participants. The context stimulus for each trial spans a pitch range of an octave, the pitch height of which is randomized independently over a two-octave range between E $\flat$ 3 and D5.<sup>3</sup> For example, for each participant, for each scale, the lowest pitch of a trial varies from E $\flat$ 3 to E $\flat$ 4 (and the highest from D4 to D5) but across all the diatonic trials the same 7 pitch classes are heard in the context. For the diatonic scale, for example, one participant might hear the pitch classes A $\flat$  B $\flat$  C D $\flat$  E $\flat$  F G and another might hear A B C $\sharp$  D E F $\sharp$  G $\sharp$ . These pitch classes are heard in one of two octaves, depending the pitch height of the octave range for the trial. The trials for each scale are kept adjacent to facilitate learning of the context scale's pitch classes.

Each pitch in the context melody plays for 200ms, with an inter-onset interval (IOI) of 250ms; all pitches have the same articulation and amplitude. Articulation involves a linear ramp up and ramp down, each of 20ms. The probe tone sounds after one second of silence, for one second. From the highest pitch of the context, to a pitch an octave lower (just below the lowest pitch of the context melody) all pitches of the equal tempered chromatic scale (numbering 13) are probed.<sup>4</sup>

The experiment also included a block of probe triads, which is discussed in Chapter 3.

### Procedure

Half the participants were asked to rate the "fit" of probe tones given the sounding of a context melody; the other half were asked to rate the "stability" of probe tones given the sounding of a context melody. In both cases, participants gave their ratings on a 7-point Likert scale: for the fit ratings, presented on-screen horizontally, the left-most point was labelled 'very bad fit', the middle point 'neither good nor bad fit' and the right-most 'very good fit'; for the stability ratings, these markers read 'very unstable', 'neither stable nor unstable' and 'very stable' respectively. Participants asked to rate stability were also informed that 'a musical sound is considered to be stable if it does not need to move (resolve) to another music sound'. Participants

<sup>3</sup>The upper bound for the overall range was erroneously stated as 'E $\flat$ 5' in the preregistration.

<sup>4</sup>The pitch an octave below the highest pitch of the context melody was included as a probe in order to test for effects of pitch height within a pitch class. Such a test however was deemed unnecessary during initial analysis.

asked to rate fit were asked simply to rate ‘how well the final musical sound fits the context melody’.

The experiment is grouped (musician or non-musician), and the block order is randomized (probe tones or probe triads first) via a controlled randomisation wherein each order is arrived at the same number of times. One factor is tested between subjects – whether the participants rate fit or stability – and another – the probe-context combination – tested within subjects.

For the probe tone block each participant heard each combination of probe and context twice. The order of probes was randomized within the contexts, whose order was also randomized. Within each trial, the order of pitches in the context was also randomized. All randomizations are independent of each other.

The probe tones block consisted of 2 iterations of 12 probe tones paired with 3 context scales to make 72 trials.

Between the two experimental blocks (a block of 72 probe tone trials and a block of 150 probe triad trials) the participants completed a survey including the Goldsmith MSI Questionnaire, in order to obtain an index for musical sophistication to be used as a variable in analysis. Additional demographic questions followed the Goldsmith MSI Questionnaire to facilitate future analysis of possible effects of enculturation (these are not analysed here).

Before the experimental trials, each block begins with 6 practice trials, leading to  $72 + 150 + 6 \times 2 = 234$  trials for the whole experiment, which took around 50mins in total.

## 2.3.2 Results

### Test for tonal hierarchies

Comparisons of Bayesian models of average ratings suggest that a tonal hierarchy is observed for all three scales. A model using just the RPC of the probe tone as a predictor performed significantly better than a model of just intercept for all three scales, as shown in Table 2.1.

TABLE 2.1: LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 1

Scale	Null – Alternative LOOIC	SE	Signif
Diatonic	584.6	50.6	Yes
Harmonic minor	538.0	48.8	Yes
Jazz minor	261.4	36.6	Yes

*SE* is the standard error of the associated LOOIC comparison. *Signif* indicates whether or not the difference in LOOIC between the null and alternative models for each scale is significant – if the difference in LOOIC is more than twice its associated SE.

### Comparison of fit and stability ratings

To give an overall picture of the difference between fit and stability the average ratings of fit and stability for all participants are shown in Fig. 2.1 (diatonic scale), Fig. 2.2 (harmonic minor scale), and Fig. 2.3 (jazz minor scale). In all plots of average ratings in this chapter, the scale-tones are numbered as RPCs – i.e., according to their distance in semitones above the theoretical tonic. Error bars are from 95% confidence intervals obtained from 1000 bootstrapped samples.

To test our hypothesis considering these differences, for each scale, ratings of fit and ratings of stability were first converted to Z-scores. Then, for each scale-probe combination, a Mann-Whitney *U*-test was run, comparing the scores for fit to those for stability. After applying Bonferroni corrections, we find that the leading tone – RPC 11 – of the diatonic scale received significantly higher fit than stability ratings ( $p = .001$ ). We find this also for the leading tone of the harmonic minor scale ( $p = .04$ ), but not of the jazz minor scale. Finally, the supertonic (second degree) of the harmonic minor scale also received significantly higher fit than stability ratings ( $p = .002$ ). Our first hypothesis is thus supported.

Though significant differences were found for particular RPCs within particular scales, we cannot so far say whether or not our Bayesian ordinal regression model benefits from the inclusion of an effect of task – whether ratings were of fit or stability. This question is explored at the beginning of the results for Model 2.1. It is found that the model without task performs significantly worse. Accordingly, in the following model we average fit and stability ratings together as ‘average ratings’.

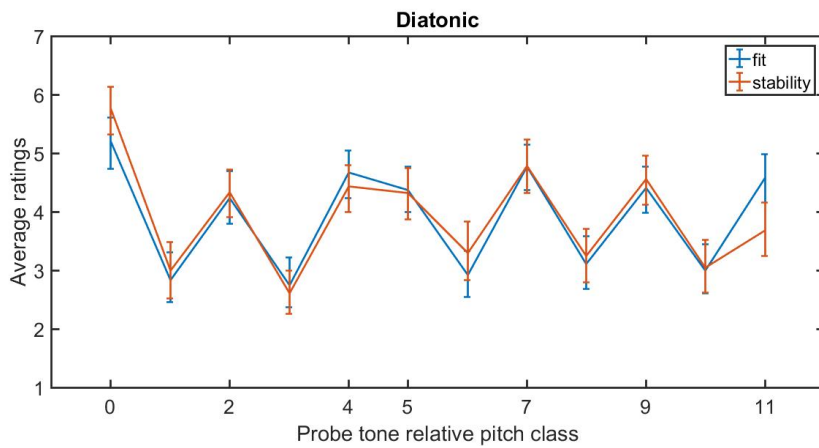


FIGURE 2.1: Average fit vs stability ratings for probe tones after the diatonic context. Scale-tones are numbered as RPCs. Error bars are from 95% confidence intervals obtained from 1000 bootstrapped samples.

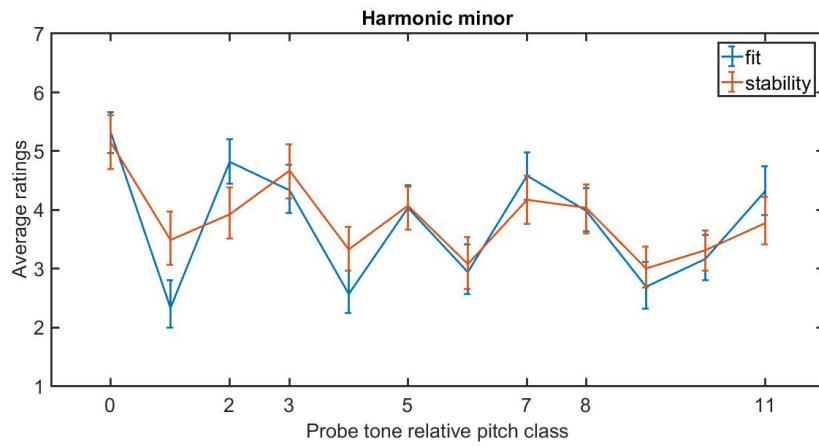


FIGURE 2.2: Average fit vs stability ratings for probe tones after the harmonic minor context

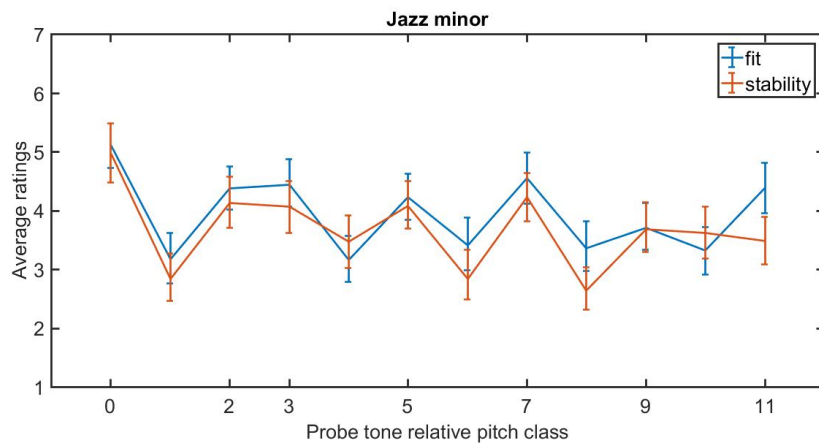


FIGURE 2.3: Average fit vs stability ratings for probe tones after the jazz minor context

### Descriptive model

Before the ordinal mixed effects models used to test the second and third hypotheses, a linear model was run comparing observed ratings for each scale-probe combination, averaged over all other variables, to SPCS's predictions.<sup>5</sup> Figures 2.4, 2.5 and 2.6 plot the SPCS prediction from this model against the average observed ratings for the diatonic, harmonic minor and jazz minor scales respectively. For all three scales we can see that one RPC – the major tonic of the diatonic scale and the (minor) tonics of the harmonic minor and jazz minor scales – was rated higher than the other RPCs included in the context (the scale-tones). For the diatonic and harmonic scales these RPCs were rated higher than all the chromatic RPCs (the pitch classes that did not appear in the context). For the jazz minor scale this was not the case, due largely to comparatively low ratings of diatonic RPCs 9 and 11 (pitch classes 9 and 11 semitones above the [theoretical minor] tonic of the scale) – and comparatively high ratings of RPC 10.

With an  $R^2$  value of .85 (adjusted, .84) SPCS models the data well. The tonics (RPCs of 0 for each scale) are notable as exceptions, along with RPC 9 of the jazz minor scale (the pitch class 'A' in the C jazz minor scale, for example), which we can see was rated lower than a SPCS model predicts.

In contrast to these simple linear models, the ordinal mixed-effects models below provide predictions for the ratings given for each individual trial.

---

<sup>5</sup>Given the ordinal nature of the dependent variable, a linear model is not ideal; however it makes for simpler interpretation for these data. This, and the fact that many previous probe-tone experiments used linear models which we wish to compare to, is why we use a linear model. Inspection of a plot of the residuals suggested that a linear model is appropriate for our data.

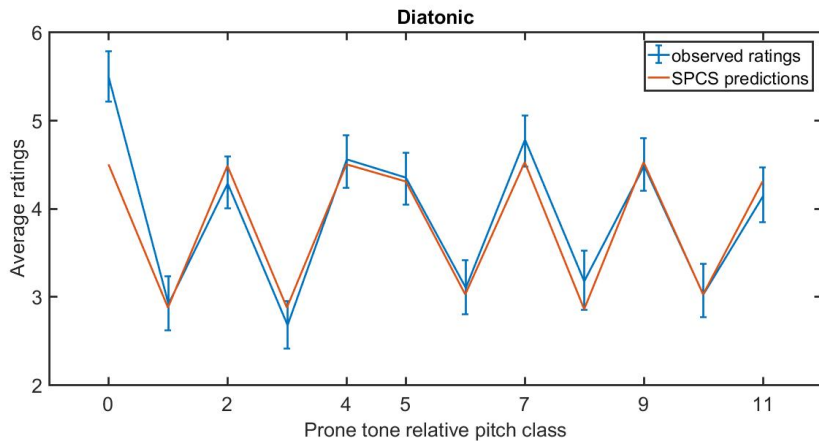


FIGURE 2.4: Average ratings for probe tones after the diatonic context compared to SPCS predictions

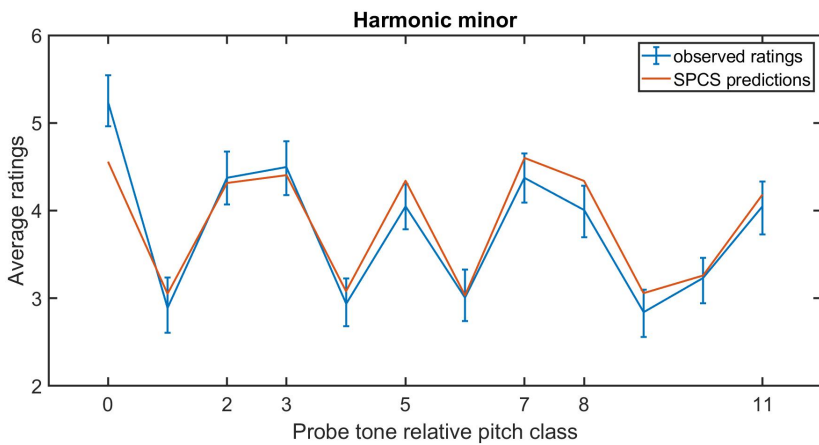


FIGURE 2.5: Average ratings for probe tones after the harmonic minor context compared to SPCS predictions

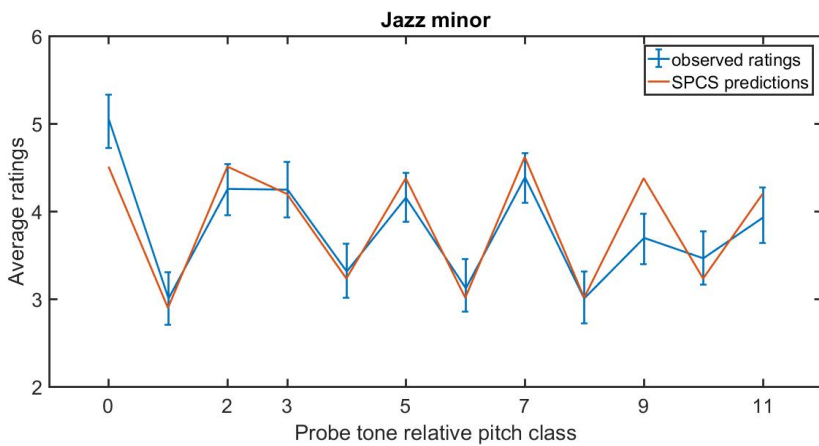


FIGURE 2.6: Average ratings for probe tones after the jazz minor context compared to SPCS predictions



**H2&3: Model 2.1**

Two alternative models were run, differing only in the presence or absence of an effect of Task (fit or stability). A LOOIC comparison of the two models favours the model without Task (Table 2.3) and it is chosen as Model 2.1 accordingly. For brevity only the significant effects are shown for this model in Table 2.2, and in the tables of later ordinal models, along with any associated conditional main effects and the intercepts. A table including all effects is included in Appendix B (Table B.2).

TABLE 2.2: Model 2.1 significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.19	0.14	-3.48	-2.91	6404	1.00
Intercept[2]	-1.45	0.13	-1.71	-1.19	6603	1.00
Intercept[3]	-0.11	0.13	-0.37	0.14	6655	1.00
Intercept[4]	0.67	0.13	0.42	0.92	6666	1.00
Intercept[5]	2.01	0.13	1.74	2.27	6795	1.00
Intercept[6]	3.68	0.15	3.39	3.97	7386	1.00
MusSoph	-0.13	0.10	-0.34	0.07	5142	1.00
Height	0.18	0.08	0.02	0.35	6875	1.00
RelHeight	0.27	0.07	0.13	0.42	9443	1.00
Height <sup>2</sup>	0.01	0.06	-0.12	0.13	8135	1.00
RelHeight <sup>2</sup>	-0.12	0.06	-0.23	-0.01	10676	1.00
Previous	0.35	0.07	0.22	0.48	8957	1.00
Recency	0.73	0.27	0.20	1.28	7871	1.00
SPCS	0.58	0.10	0.39	0.77	7803	1.00
Prevalence	0.52	0.09	0.34	0.70	8553	1.00
TrialNo	-0.01	0.07	-0.15	0.12	9806	1.00
MusSoph:Previous	-0.11	0.05	-0.22	-0.00	7853	1.00
MusSoph:Recency	0.66	0.23	0.22	1.14	8186	1.00
MusSoph:SPCS	0.26	0.09	0.09	0.42	8752	1.00
MusSoph:Prevalence	0.28	0.08	0.13	0.43	8281	1.00
Height <sup>2</sup> :TrialNo	0.07	0.04	0.00	0.14	12299	1.00

Significant population-level effects for Model 2.1 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

The six intercepts represent 'cutpoints' between the 7 ordinal values of ratings;

the Intercept values are of a latent continuous variable; the Estimate value corresponds to the change in this latent variable that is associated with an increase of 1 standard deviation in the value of the effect for continuous variables, or with an increase from a value of 0 to a value of 1 for binary variables (Recency only, in Table 2.2). For example, given the Estimate value for SPCS of 0.58 a change in SPCS of 2 SD would take a rating of 4 (an Estimate between  $-0.11$  and  $0.67$ ), for example, to a rating of 5 (Estimate between  $0.67$  and  $2.01$ ). We interpret effects with a magnitude of about 0.2 to be small, those of about 0.5 to be medium, those of about 0.8 to be large.

Comparing this model's LOOIC to that of its associated reduced model (equivalent but for the absence of SPCS and Prevalence), we find that Model 2.1 performs significantly better (see Table 2.3 for details), confirming H1.

Significant as conditional main effects in this model are SPCS (medium effect size, evid. ratio  $> 11999$ ), Prevalence (medium, evid. ratio  $> 11999$ ), Recency (large, evid. ratio 254.32), Previous (medium, evid. ratio  $> 11999$ ), Height (small, evid. ratio 65.3) and RelHeight (small, evid. ratio 11999) in the positive direction and RelHeight<sup>2</sup> in the negative direction (small, evid. ratio 63.17).

Significant interaction effects with MusSoph are with Recency (large, evid. ratio 479), SPCS (small, evid. ratio 443.44), and Prevalence (small, evid. ratio 3999) in the positive direction and Previous (small, evid. ratio 49) in the negative direction. An interaction effect between Height<sup>2</sup> and TrialNo, though rather small, is also significant (evid. ratio 43.28), in the positive direction. None of MusSoph, TrialNo or Height<sup>2</sup> are significant as conditional main effects.

We can interpret from the interactions with MusSoph that the more musically sophisticated participants were better at using cues from the stimulus to shape their ratings and that they were also less likely to be influenced by their previous rating.

The influence of Height and Height<sup>2</sup> together on ratings results in an inverted U-shaped curve for the earlier trials and a linear curve for later trials. For a plot depicting this relationship, see Figure B.1 in Appendix B. We can see from this that there exists at the start of the experiment a negative effect of Height<sup>2</sup> that is absent through the middle and later parts of the experiment.

Finally, the conditional effect of RelHeight, shown in Appendix B (Figure B.2), shows an inverted U-shape with a peak towards higher values of RelHeight, reflecting the positive effect of RelHeight on ratings along with the negative effect of RelHeight<sup>2</sup>.

The *ppcheck* plot for this model, shown in Figure 2.7, confirms the validity of the model. The pseudo- $R^2$  value of the model is .57. The first number in each pair refers to the scale, where ‘1’ is the diatonic, ‘2’ is the harmonic minor and ‘3’ is the jazz minor; the second number in each pair refers to the RPC of the probe. ‘*y*’ represents the data and ‘*yrep*’ represents the distribution of predictions obtained from random samples of parameter values from the posterior predictive distribution for Model 2.1.

### Exploratory analysis

The effects of SPCS and Prevalence are intended to account for two competing theories for the cognition of harmonic tonality. Prevalence models a top down process based on long term statistical learning of the frequency of occurrence of RPCs in tonal-harmonic music whereas SPCS is a bottom-up process based on a psychoacoustic response to the frequency content of the stimulus. SPCS may be able to account for Prevalence, where the statistics of music learned may be themselves shaped by psychoacoustic features. We explore the relationship between SPCS and Prevalence by running two models, equivalent to Model 2.1, but with Prevalence removed, or with SPCS removed, respectively. With Prevalence removed, SPCS has an effect size of 0.89, which is larger than the effect sizes of SPCS or Prevalence in Model 2.1, but smaller than their sum (the model is shown in Appendix B in Table B.3). With SPCS removed, Prevalence has an effect size of 0.56, which is approximately the same as its effect size in Model 2.1. No significant difference in performance was found between these two models, both of which were significantly outperformed by Model 2.1 (see Table 2.3 for details).

TABLE 2.3: LOOIC comparisons for Experiment 1 Bayesian ordinal mixed effects models

Model compared to Model 2.1	Model – Model 2.1 LOOIC	SE	Signif
Model 2.1 + Task	14.4	4.6	Yes
Model 2.1 – SPCS – Prevalence	1173.9	78.9	Yes
Model 2.1 – Prevalence	351.6	36.6	Yes
Model 2.1 – SPCS	317.2	41.8	Yes
Model 2.1 – SPCS + ScaleTone	2.7	14.2	No
Model compared to Model 2.1 – Prevalence	Model – (Model 2.1 – Prevalence) LOOIC	SE	Signif
Model 2.1 – Prevalence – SPCS + ScaleTone	128.6	20.0	Yes
Model 2.1 – SPCS	–64.6	62.4	No

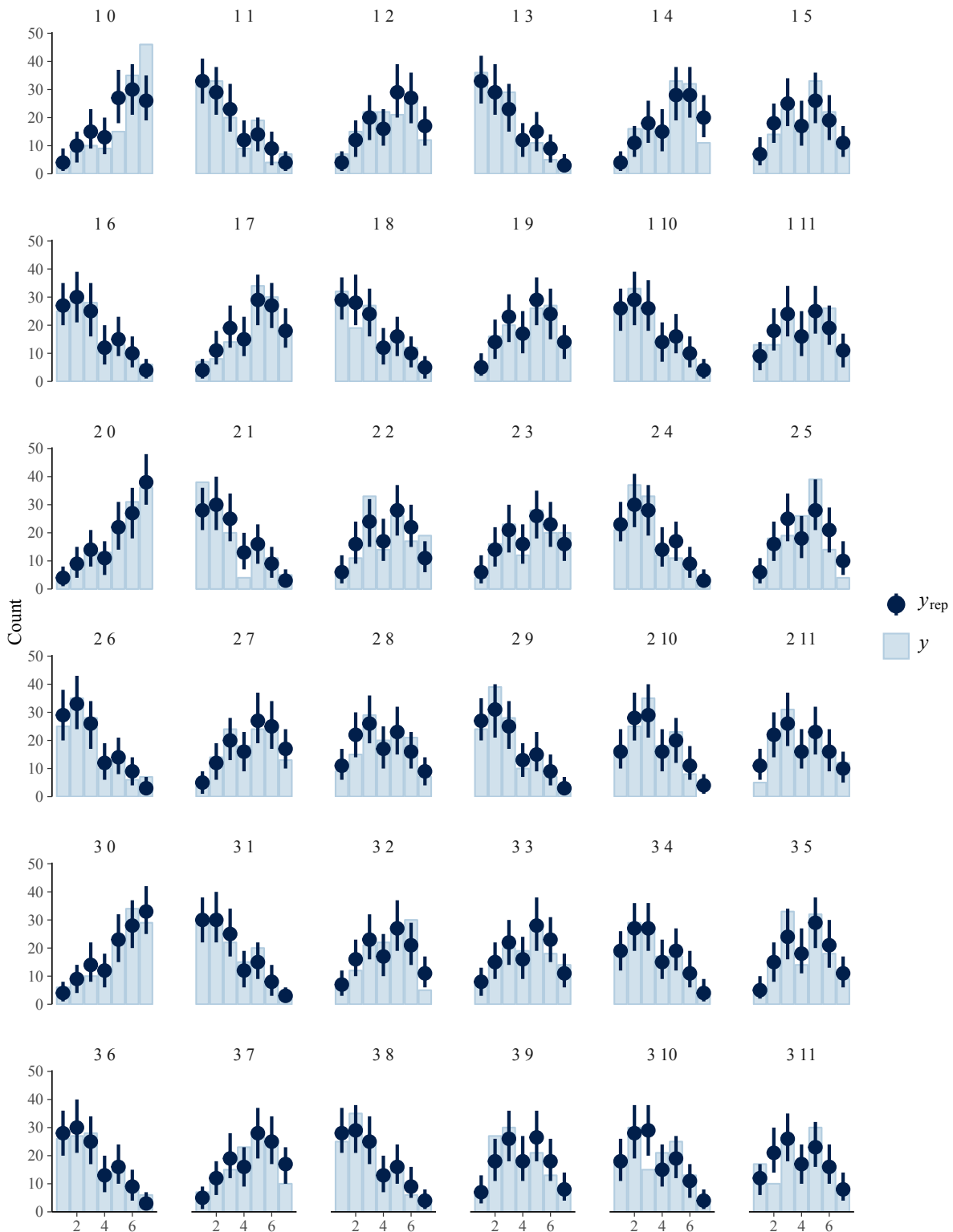


FIGURE 2.7: *ppcheck* – posterior predictive check – grouped by scale-probe combination. Each plot shows the number of each of the 7 Likert scale ratings (from ‘very unstable’ or ‘very bad fit’ on the left to ‘very stable’ or ‘very good fit’ on the right) for each probe after each scale observed ‘ $y$ ’ and predicted by the model ‘ $y_{rep}$ ’. The first number in the title for each plot refers to the scale, where ‘1’ is the diatonic, ‘2’ is the harmonic minor and ‘3’ is the jazz minor; the second number in each pair refers to the RPC of the probe. ‘ $y$ ’ represents the data and ‘ $y_{rep}$ ’ represents the distribution of predictions obtained from random samples of parameter values from the posterior predictive distribution for Model 2.1.

We might also ask what information SPCS can provide beyond simply a binary prediction based on the partition of probes between pitch classes played in the context melodies, and pitch classes that are not. From our descriptive model we know that there is more complexity in ratings than can be explained by this binary partition: In each scale one RPC (the tonic) was rated significantly higher than the others. We can see that changing a single pitch class in the context from a diatonic to a harmonic minor scale (in C major / A minor for example, the diatonic scale consists of the pitch classes C, D, E, F, G, A and B, and the harmonic minor of the pitch classes A, B, C, D, E, F, G#), for example, influences not only the ratings of pitch classes that shifted from being scale-tones to chromatic RPCs and vice versa, but also of the tonic RPCs. Though the tonics are not predicted by SPCS, one can see that within the non-tonic scale-tones and within the non-scale-tones there are differences in SPCS which do appear to be reflected in the ratings.

We define the additional predictor *ScaleTone*, coded as 1 when the probe is a scale-tone – i.e., a pitch class included in the context stimulus – or as 0 otherwise. We define a model identical to Model 2.1 but for the inclusion of *ScaleTone* as an effect instead of SPCS. In this model, *ScaleTone* and the interaction between *ScaleTone* and *MusSoph* are both significant in the positive direction; a LOOIC comparison between the models suggests no difference in out-of-sample predictive performance between the two (details shown in Table 2.3).

In a model with *ScaleTone* and without SPCS or *Prevalence*, *ScaleTone* and *MusSoph:ScaleTone* are both positive and significant. A positive interaction between *ScaleTone* and *InContTrialNo* also reaches significance in this model. This model is found to be significantly worse (see Table 2.3) than a model equivalent to this one but with SPCS instead of *ScaleTone*.

The significant population-level effects and associated conditional effects for each of these models are included in Appendix B (Tables B.4 and B.5).

We also considered more complex models of *Recency* and *Primacy*, namely the pitch distance of the probe in semitones from the final and initial pitches of the context melody, respectively, and the square of these values. We found that these did not improve the model as judged by LOOIC and so they are not used in any of the models detailed in this dissertation.

### 2.3.3 Discussion

Considering our exploratory analyses, a model equivalent to Model 2.1 but with ScaleTone replacing SPCS is judged via LOOIC to be insignificantly different in performance to Model 2.1, but if Prevalence is removed from both models the model with SPCS is judged to perform significantly better. This suggests that SPCS is a better measure than ScaleTone of a mechanism at least partially accounted for by Prevalence (which we can safely assume is related to long-term statistical learning) as well as of another mechanism that is not predicted by Prevalence. The *ppcheck* plot (Figure 2.7), however, suggests that neither Prevalence nor SPCS are able to account for the high rating of the major tonic of the diatonic scale, where SPCS alone predicts the major tonic to have equally high ratings as RPCs 2, 4, 7 and 9, and Prevalence alone predicts that it should receive the second highest rating, with RPC 7 (the dominant), the highest. Perhaps there is an effect of learning, which we assume may be stronger for musicians, that is not being accounted for by either SPCS or Prevalence: Perhaps participants are able to recognise the scale's conventional tonics, and rate them highly accordingly.

We have discussed the difficulty in appropriately using prevalence for our experimental design. While possible for these major and minor scales, this will not be possible for less common scales for which we can assume possible effects of familiarity will be substantially diminished. If there are clearly hierarchies of perceived fit or stability, and we find SPCS to contribute to a model for goodness-of-fit or stability ratings for tones and after the context of less common scales, then this strengthens the argument for a psychoacoustic description of the cognition of harmonic tonality. Experiment 2 involves such a set of scales as stimulus.

## 2.4 Experiment 2: Less familiar scales

In Experiment 1, the common usage of the scales used for context meant that the effect of statistical learning needed to be considered in relation to the perceived fit/stability of probes sounded after the context scales. A second experiment was devised using as context less familiar scales for which this effect should be reduced, namely: hexatonic, octatonic, harmonic major and double harmonic scales (detailed below). This reduction should aid in a consideration of the effect of SPCS on tonal fit. The relatively common (though arguably still less common in tonal-harmonic

music than the three scales tested in Experiment 1) pentatonic and blues scales were also included given the room for two extra scales to be tested within the experimental time frame. The scales used for stimuli in this experiment, notated in common modes and beginning on C for convenience, are:

1. Pentatonic: C E $\flat$  F G B $\flat$
2. Blues: C E $\flat$  F F $\sharp$  G B $\flat$
3. Hexatonic: C E $\flat$  E G A $\flat$  B
4. Octatonic: C D E $\flat$  E F $\sharp$  G A B $\flat$
5. Harmonic major: C D E F G A $\flat$  B
6. Double Harmonic: C D $\flat$  E F G A $\flat$  B

All these were tuned to 12-tone equal temperament.

If the results can be modelled accurately by SPCS a psychoacoustic description of the cognition of tonality is supported. Participants, including both musicians and non-musicians also completed the Goldsmith MSI Questionnaire in order to account for effects of musical sophistication upon their responses.

### 2.4.1 Method

The stimulus is as in Experiment 1, but for

1. The context scales: pentatonic, blues, hexatonic, harmonic major, double harmonic and octatonic, tuned to 12-tone Equal Temperament.
2. The probe triads (discussed in Chapter 3).

The experimental procedure differs in the number of trials in each block: The probe tone block includes  $12 \times 6 \times 2 = 144$  trials; the probe triad block includes  $(6 + 5) \times 7 = 77$  trials (not discussed in this paper). Before the experimental trials, each block begins with 6 practice trials, leading to  $72 + 150 + 6 \times 2 = 233$  trials for the whole experiment.

## Participants

Thirty-two musicians (participants who reported having received 5 or more years of music experience) and 32 non-musicians were recruited for the experiment. Non-musician participants were first-year university students recruited through Western Sydney University School of Psychology and Social Science's SONA system and received credit points towards their degrees for their participation. Musician participants were recruited via personal connection and received a \$30 reimbursement for their time and travel to the university campus. All participants reported normal hearing capabilities. Survey data were received for only 28 musicians and 29 non-musicians. Of these, five (musician) participants reported having received 10 or more years of musical training and one reported having absolute pitch. Participants had a mean age of 21.5 years, with a SD of 4.0 years. Of the 28 musicians 16 were female, and out of the 29 non-musicians 24 were female. This research was approved by the Western Sydney University Human Research Ethics Committee under the number H11908.

## 2.4.2 Results

### Test for tonal hierarchies

For all scales apart from the octatonic scale, a model using just the RPC of the probe tone as a predictor of average ratings performed significantly better than a model of just intercept, as shown in Table 2.4. This shows strong evidence that there are tonal hierarchies for all scales except the octatonic, for which the evidence is not significant. We should note here however that the tonal hierarchy (if we can, in this instance, call it that) for the hexatonic scale (as well as the octatonic scale) has only two levels – scale-tone and non-scale-tone – as we see below in Fig. 2.12 (and Fig. 2.13).



TABLE 2.4: LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 2

Scale	Null - Alternative LOOIC	SE	Signif
Harmonic major	103.0	28.2	Yes
Double harmonic	96.4	31.4	Yes
Pentatonic	340.6	46.6	Yes
Hexatonic	90.2	22.6	Yes
Octatonic	24.6	18.6	No
Blues	124.0	31.6	Yes

### Comparison of fit and stability

Plotting average ratings for fit and stability for these scales suggests no significant differences between fit and stability ratings for any RPCs of any scale for participants overall (or for the musicians or non-musician participants only). A Mann-Whitney  $U$ -test for fit vs stability for all 12 RPCs of all six scales indeed returns no significant differences. Our first hypothesis thus is supported. In the descriptive model below, fit and stability ratings are averaged together as ‘average’ ratings, and task is not included as an effect in Model 2.2, which is used to test our second and third hypotheses.

### Descriptive model

A linear model was run with SPCS as the sole predictor of ratings averaged over participants, task and trials.<sup>6</sup> Figures 2.8, 2.9 2.10, 2.11, 2.12, and 2.13 show SPCS’s predictions for the six scales of Experiment 2, against the average observed ratings, with 95% confidence intervals from 1000 bootstrapped samples.

The most obvious feature of many of these plots in comparison to Experiment 1 is the lack of a single distinct tonic (a single pitch rated clearly higher than any others), with the pentatonic and blues scales (which are arguably the most common) the only exceptions. The pitch which would be the tonic of the major mode (arguably the most common mode) of the pentatonic scale did receive higher ratings. The same is true for the familiar tonic of the blues scale. Neither tonic is predicted by SPCS (the single distinct tonics of all three scales in Experiment 1 were also not predicted by SPCS), where we can see that SPCS predictions are lower than observed ratings. The blues scale can be thought of as the pentatonic scale with an RPC added between

<sup>6</sup>As in Experiment 1, inspection of a plot of the residuals suggested that a linear model is appropriate for the averaged data.

RPCs 5 and 7. For this scale, along with RPC 5, the RPCs that correspond to the familiar major and minor tonics of the pentatonic scale – RPCs 0 and 3 respectively – were rated higher than the non-scale-tones. In the harmonic major scale, along with RPC 2 and 4, the two RPCs that obtained the (equal) highest predicted ratings – RPC 0 and RPC 7 – received higher ratings than the four RPCs that obtained the (equal) lowest predicted ratings – RPCs 1, 3, 6 and 10.

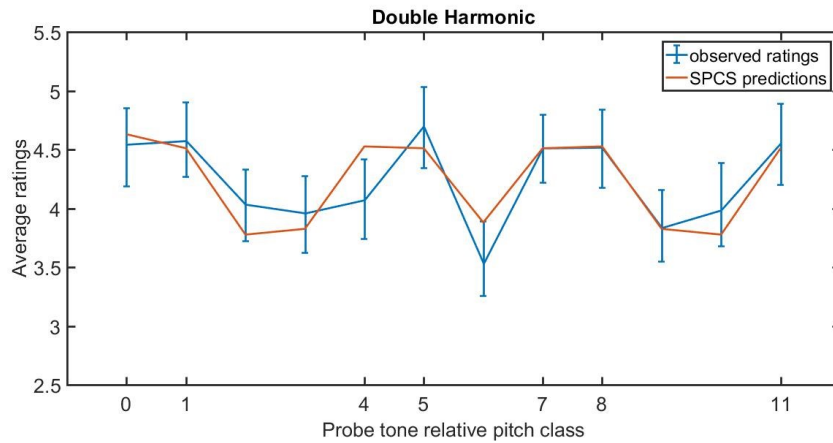


FIGURE 2.9: Average ratings for probe tones after the double harmonic context compared to SPCS predictions

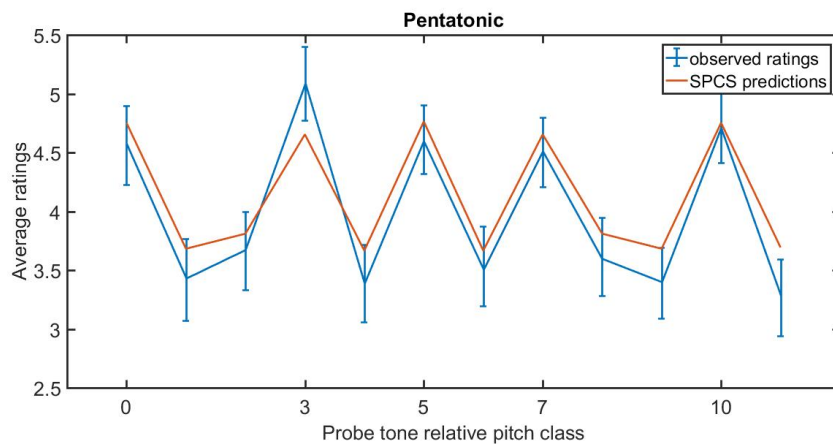


FIGURE 2.10: Average ratings for probe tones after the pentatonic context compared to SPCS predictions

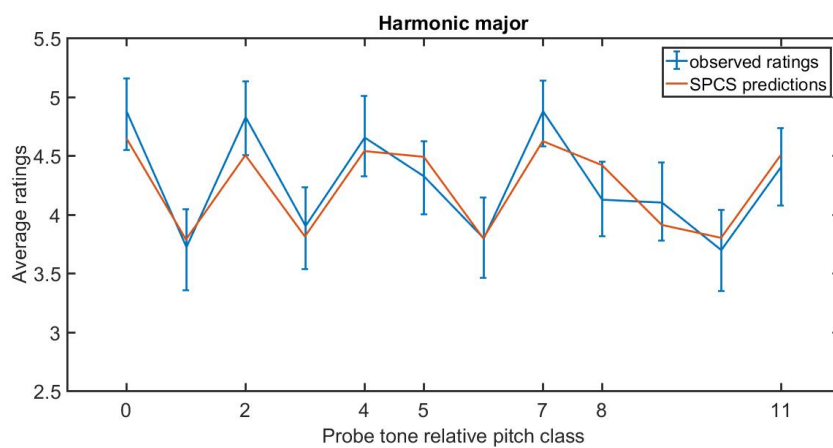


FIGURE 2.8: Average ratings for probe tones after the harmonic major context compared to SPCS predictions

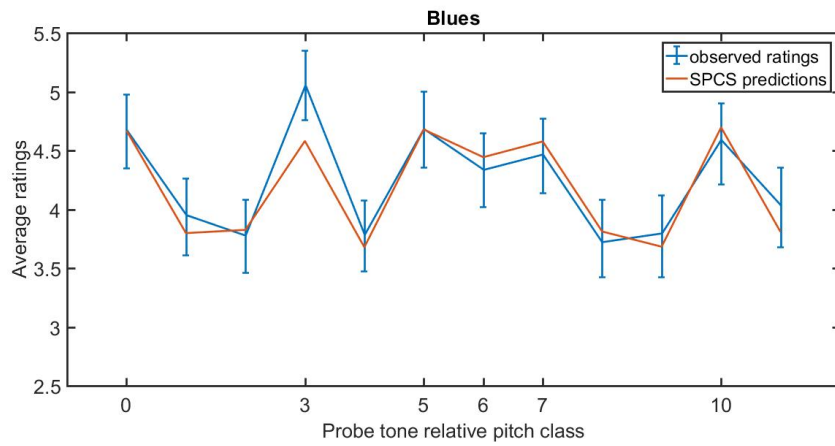


FIGURE 2.11: Average ratings for probe tones after the blues context compared to SPCS predictions

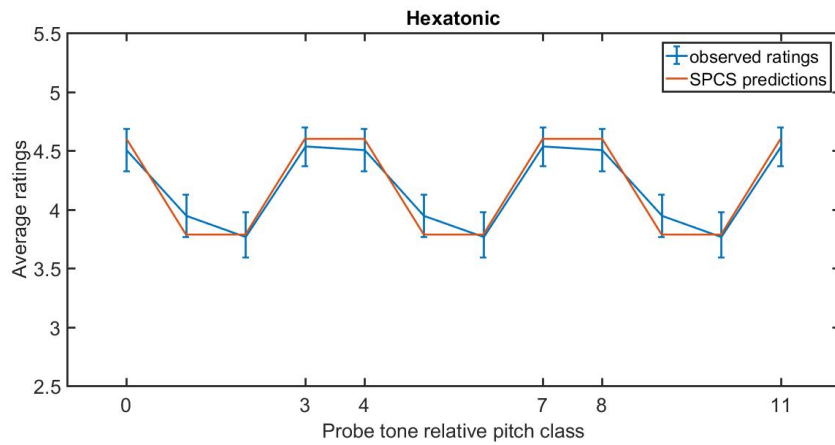


FIGURE 2.12: Average ratings for probe tones after the hexatonic context compared to SPCS predictions

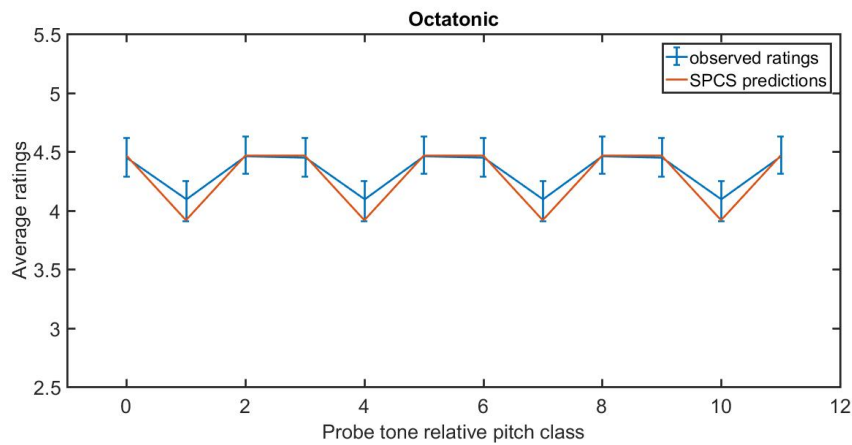


FIGURE 2.13: Average ratings for probe tones after the octatonic context compared to SPCS predictions

Error bars in the plots suggest that only for the pentatonic, hexatonic and octatonic scales were all scale-tones rated significantly higher than all non-scale tones. Since the error bars in the plots do not assess the significance of comparisons between probes, however, this can only be suggested. The data obtained for the hexatonic and octatonic scales benefit from a combination of these scales' structure and the stimulus presentation used in the experiment: Because the RPCs of the lowest and highest pitches are randomized, and the structure of the scale repeats three (hexatonic) or four (octatonic) times in an octave, these scales effectively span only four semitones (hexatonic), and three semitones (octatonic); so, for each RPC there are three times (hexatonic) or four times (octatonic) as many observations for each RPC as for the other scales.

One can observe that for the octatonic and hexatonic scales SPCS is equivalent to our binary ScaleTone effect introduced in Section 2.2.3 above. Of all scales the range of average ratings covered by the relative pitch classes is smallest for the octatonic scale, next smallest for the hexatonic, and largest for the pentatonic.

Overall SPCS was a very strong predictor though perhaps slightly weaker for this set of scales than for the more familiar set of Experiment 1:  $R^2 = .83$ , adjusted .83 for Experiment 2, compared to  $R^2 = .85$ , adjusted .84 for Experiment 1.

### **H2&3: Model 2.2**

In Experiment 1 the stimulus included only 7-note scales, whereas in Experiment 2 scales of 5, 6 and 8 notes are also included. Accordingly, in this model an effect of ScaleSize is included (standardized), along with its interaction with MusSoph, TrialNo, TrialNo<sup>2</sup>, BlockOrder and InContTrialNo. For the four musician and two non-musician participants whose survey data were lost or not collected we imputed Musical Sophistication scores equal to the mean for their groups – either musician or non-musician. Listed for Model 2.2 (Table 2.5) are all the effects that emerge as significant from such a model, along with the intercepts and associated conditional effects. A table including all effects is shown in Appendix B (Table B.6).

TABLE 2.5: Model 2.2 Significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.27	0.11	-3.49	-3.05	5216	1.00
Intercept[2]	-2.09	0.10	-2.29	-1.88	5043	1.00
Intercept[3]	-0.93	0.10	-1.13	-0.73	5014	1.00
Intercept[4]	-0.16	0.10	-0.36	0.04	5052	1.00
Intercept[5]	0.99	0.10	0.79	1.20	5100	1.00
Intercept[6]	2.44	0.11	2.23	2.65	5455	1.00
MusSoph	-0.16	0.10	-0.35	0.03	4297	1.00
Height	0.13	0.08	-0.03	0.28	4850	1.00
RelHeight	0.34	0.08	0.18	0.49	4700	1.00
Height <sup>2</sup>	0.06	0.04	-0.02	0.15	7863	1.00
RelHeight <sup>2</sup>	-0.31	0.06	-0.43	-0.20	6595	1.00
Primacy	-0.24	0.12	-0.48	-0.01	13697	1.00
Previous	0.33	0.07	0.19	0.47	5923	1.00
Recency	0.54	0.14	0.26	0.82	11496	1.00
SPCS	0.48	0.09	0.30	0.65	5687	1.00
MelCont	-0.04	0.03	-0.11	0.02	16017	1.00
TrialNo	0.05	0.05	-0.04	0.15	11322	1.00
InContTrialNo	-0.05	0.04	-0.12	0.02	16338	1.00
ScaleSize	-0.02	0.05	-0.11	0.06	11648	1.00
MusSoph:Recency	0.32	0.12	0.09	0.57	11244	1.00
MusSoph:SPCS	0.29	0.09	0.11	0.46	5014	1.00
MelCont:TrialNo	-0.06	0.02	-0.11	-0.02	17790	1.00
Height:InContTrialNo	-0.07	0.03	-0.13	-0.00	13908	1.00
Height <sup>2</sup> :ScaleSize	0.05	0.02	0.00	0.10	15589	1.00
SPCS:ScaleSize	-0.09	0.03	-0.15	-0.04	16106	1.00

Significant population-level effects for Model 2.2 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

As in Model 2.1, SPCS (evid. ratio > 11999), Recency (evid. ratio > 11999), Previous (evid. ratio > 11999) and RelHeight (evid. ratio 11999) are significant as positive main effects and RelHeight<sup>2</sup> as a negative main effect (evid. ratio > 11999). The effect of Recency in this model is of medium, rather than large size. SPCS remains medium and the others small. Unlike Model 2.1, Height is not significant as a main effect (evid. ratio 16.86). Primacy is also significant as a small main effect in this model, in the negative direction however (evid. ratio 52.1). We cannot think of an explanation for this.

A conditional effects plot of RelHeight, shown in Appendix B (Figure B.3), reveals a very similar relationship to the one found in Model 2.1. Prevalence, of course, does not appear in this model.

In terms of interactions, though interactions of MusSoph with SPCS (evid. ratio 1713.29) and Recency (evid. ratio 229.77) are again significant in the positive direction, the interaction of MusSoph with Previous is not significant in this model (evid. ratio 9.76). Both interactions are of small size. With the addition of ScaleSize as a variable in this model, we see significant interaction effects of ScaleSize with Height<sup>2</sup> (positive, evid. ratio 54.81) and SPCS (negative, evid. ratio 2999) where for scales of fewer pitch classes ratings are less affected by Height<sup>2</sup>, and more by SPCS. The effect of this interaction is most easily understood by examining the conditional effects plot of Height:ScaleSize, shown in Appendix B (Figure B.4). In this model the interaction between Height and InContTrialNo is also significant, in the negative direction (evid. ratio 56.97), suggesting that as a particular context scale becomes more familiar to participants their ratings are less affected by pitch height. Finally a significant negative interaction between MelCont and TrialNo (evid. ratio 254.32) suggests that participants are less influenced by melodic continuity for later experimental trials. Where significant interactions with MusSoph involve larger effect sizes than other significant interactions it may be suggested that the musical sophistication of the participants is most influential to the effect of other predictors.

Significant interaction effects with MusSoph are present as expected. A null model was run, which was identical but for the absence of effects of SPCS. The difference in LOOIC between Model 2 and its associated null model is 805.6, in favour of Model 2., with a standard error of 76.5 (included in Table 2.7.) This confirms our hypothesis that fit and stability ratings can be predicted by SPCS.

The pseudo- $R^2$ , at .51, is lower for this model than for Model 2.1.

### **Combined analysis: Model 2.3**

The data from Experiments 1 and 2 can be combined and a model equivalent to Model 2.2 run for this combined data set. This model – Model 2.3, whose results are summarized in Table 2.6 below – is more informative than those above as it includes the most wide-ranging data and number of observations. Table B.7 in Appendix B shows all population-level effects for Model 2.3.

TABLE 2.6: Model 2.3 significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.09	0.08	-3.25	-2.93	4931	1.00
Intercept[2]	-1.75	0.08	-1.90	-1.59	4709	1.00
Intercept[3]	-0.57	0.08	-0.72	-0.42	4688	1.00
Intercept[4]	0.18	0.08	0.03	0.33	4668	1.00
Intercept[5]	1.34	0.08	1.19	1.49	4750	1.00
Intercept[6]	2.78	0.08	2.62	2.94	5121	1.00
MusSoph	-0.16	0.07	-0.29	-0.02	3707	1.00
Height	0.13	0.05	0.03	0.23	6464	1.00
RelHeight	0.28	0.05	0.18	0.38	5122	1.00
RelHeight <sup>2</sup>	-0.22	0.04	-0.30	-0.15	7124	1.00
Previous	0.33	0.04	0.25	0.42	6601	1.00
Recency	0.65	0.13	0.40	0.89	7955	1.00
SPCS	0.64	0.06	0.51	0.77	4661	1.00
MelCont	-0.03	0.03	-0.08	0.03	11662	1.00
Count	0.01	0.05	-0.08	0.10	8982	1.00
TrialNo	0.07	0.04	-0.02	0.15	7835	1.00
InContTrialNo	-0.03	0.03	-0.08	0.03	14086	1.00
ScaleSize	-0.04	0.04	-0.12	0.04	9675	1.00
MusSoph:Previous	-0.10	0.04	-0.18	-0.03	6163	1.00
MusSoph:Recency	0.43	0.11	0.21	0.65	8003	1.00
MusSoph:SPCS	0.36	0.06	0.24	0.48	5045	1.00
SPCS:TrialNo	-0.06	0.02	-0.10	-0.01	12072	1.00
MelCont:TrialNo	-0.06	0.02	-0.11	-0.02	14309	1.00
Height:InContTrialNo	-0.06	0.02	-0.11	-0.01	14024	1.00
SPCS:ScaleSize	-0.07	0.02	-0.12	-0.02	12830	1.00

Significant population-level effects for Model 2.3 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

All effects significant in Model 2.2 are significant again in this model apart from Primacy, along with all comparable effects significant in Model 2.1, apart from MusSoph:Prevalence (as Prevalence could not be included in this model) and the interaction between Height<sup>2</sup> and TrialNo. In this model MusSoph is significant also as a (small) main effect in the negative direction, suggesting that more musically sophisticated participants gave lower ratings overall.

Additionally, an interaction of SPCS with TrialNo was significant in the negative direction in this model, suggesting that the influence of SPCS is stronger towards the



beginning of the block. We are not surprised to see more effects come into significance in this model, since we have more observations than in the previous models.

This model has a pseudo- $R^2$  of .50, just below the value for Model 2.2.

### Exploratory analysis

The results of the following model comparisons are detailed in Table ???. Comparing Model 2.2 to the same model but with an added effect of Task, we find Model 2.2 to perform insignificantly better (difference in LOOIC of 8.81 with a SE of 4.80). Though not significantly better than with Task, our use of the model without Task for Model 2.2 is still beneficial as it is simpler and more generalizable, as it may be used to predict ratings whether they are of fit or of stability.

An alternative model identical to Model 2.2 but with ScaleTone replacing SPCS was run, and compared to Model 2.2. Model 2.2 was seen to significantly outperform this alternative model (difference in LOOIC of 55.22, with a SE of 15.11).

The same was done for Model 2.3, with Model 2.3 significantly outperforming its ‘ScaleTone’ alternative (difference in LOOIC of 180.90, for a SE of 21.61). ScaleTone and MusSoph:ScaleTone are significant positive effects in both these models. These results reflect those of the exploratory analyses of Experiment 1 when Prevalence is not included.

TABLE 2.7: LOOIC comparisons for Experiment 1 Bayesian ordinal mixed effects models. A significant negative  $\Delta$ LOOIC supports the model shown in the first column; a significant positive  $\Delta$ LOOIC supports the comparison model shown in the header.

<b>Model compared to Model 2.2</b>	<b>Model – Model 2.2 LOOIC</b>	<b>SE</b>	<b>Signif</b>
Model 2 + Task	8.8	4.8	No
Model 2 – SPCS	805.6	76.5	Yes
Model 2 – SPCS + ScaleTone	55.22	15.11	Yes
<b>Model compared to Model 2.3</b>	<b>Model – Model 2.3 LOOIC</b>	<b>SE</b>	<b>Signif</b>
Model 3 – SPCS + ScaleTone	180.9	21.6	Yes

### 2.4.3 Discussion

A hierarchy of perceived stability was observed in the responses for most of the scales. Stability responses were modelled significantly better with SPCS than with ScaleTone. This suggests that a tonal hierarchy emerges from the RPCs of the scales, and that this hierarchy may be predicted by SPCS.

Regarding the average ratings data discussed in Section 2.4.2, we noted that no scale showed a single clear ‘tonic’. This was predicted by SPCS (SPCS also predicted no single tonic for the scales tested Experiment 1, though results did show single clear perceptual tonics). Tonal-harmonic music, however, has been written in some of these scales, employing the use of a pitch class as a tonic. We might ask whether a perceptual tonic can be achieved even if the scale and tuning system provides no probe tone clearly more stable than all others. This is a separate empirical question, but the literature review suggests that through the contribution of other elements of music besides RPC (such as metre and the use of triads) a tonic can be achieved. In this research we do not suppose that it is necessary for a single tonic to emerge from the scale without contribution of other musical elements, only that some form of hierarchy emerges.

For this less familiar set of scales, there were no task effects significant in any models and the removal of these effects did not weaken the models. On top of this, no significant difference between fit and stability ratings was found for any RPC of any scale. This suggests that the music-theoretically appropriate differences between fit and stability that were observed in the diatonic and harmonic minor scales of Experiment 1 might be due to implicit or explicit (through music education) learning of the use of those scales. We might assume then that for arbitrary scales fit and stability do not functionally differ.

As in Experiment 1, the performance of Model 2.2 as significantly better than its associated null confirms H2 (that perceived goodness-of-fit and perceived stability of probe tones may be modelled by the SPCS of the aggregated pitches of the context and the probe pitch, and the statistical prevalence in Western music of the probe within the context scale.). H3 is confirmed by the significance in Model 2.2 of interactions with MusSoph. Models 2.2 and 2.3 are significantly weakened by the removal of effects of SPCS. We have very strong evidence that SPCS measures an effect important in the cognition of harmonic tonality and that this effect is enhanced by musical sophistication, possibly due to training of audition skills.

## 2.5 Conclusion

In Experiment 1 participants were played randomly generated, isochronous melodies using notes from the diatonic, harmonic minor and jazz minor scales. Half the participants rated the goodness-of-fit of probe tones played after these scales, and the other half rated their stability. We found that, overall, perceived goodness-of-fit can be considered equivalent to perceived stability. Having hypothesized (H1) that for specific pitch-classes ratings of fit would differ significantly from ratings of stability, we found that the leading tones of the diatonic and harmonic minor scales, as well as the supertonic of the harmonic minor scale received significantly higher fit than stability ratings. Our hypothesis was thus supported. Tonal hierarchies were statistically significant in all three scales.

Experiment 2 differed from Experiment 1 in the scales used for context melodies. Context melodies comprised notes of the pentatonic and blues scales, as well as the less familiar hexatonic, octatonic, harmonic major and double harmonic scales. For this experiment we hypothesized (H1) that given the lower level of familiarity with the scales used for the context stimulus, we would not see significant difference between fit and stability ratings for any pitch-classes. Our results supported this hypothesis. Tonal hierarchies were present in all scales apart from the octatonic scales, though for the hexatonic scale the hierarchy comprised only two levels – scale-tone and non-scale-tone.

For our first data set of familiar scales (Experiment 1), our second data set of (overall) less familiar scales (Experiment 2), and for our combined data set, we have shown that SPCS – significant as a main effect and as an interaction effect with MusSoph in our models – plays a measureable and predictable role in the cognition of tonality. Thus, our second and third hypotheses are confirmed for Experiments 1 and 2. Whether or not the probe pitch matched that of the last note of context (Recency), the rating given to the previous trial (Previous) and the pitch height of the probe relative to the lowest pitch height of the context stimulus (RelHeight) and the square of this value were also significant as main effects in all models.

Exploratory analyses demonstrated that SPCS outperforms ScaleTone (whether or not the probe is a scale-tone, i.e. whether or not it is in the context), when no effect of Prevalence is included. From this we deduce that SPCS predicts more structure in the perception of tonal fit than simply the fact that probes of RPCs heard in the context stimulus receive higher ratings than probes of RPCs not heard in the context.

We can also be confident that this cannot be completely accounted for by Prevalence.

Though we used largely unfamiliar scales in Experiment 2, they nonetheless do occur in Western common-practice music, so we cannot rule out effects of familiarity in ratings, even if we cannot feasibly test for it. However, considering the existence of cases where models of prevalence cannot easily be applied – for example, using scales for which there is no associated corpus data – we can already be sure that SPCS is a more widely applicable model than Prevalence: unlike Prevalence and top down models, SPCS can be immediately generalized to novel stimuli. It would seem wise to use truly novel (microtonal) scales for our context stimulus in future work both to further diminish effects of familiarity, and to test generalization. Such an experiment is discussed in Chapter 6.

## Chapter 3

# Experiments 1 & 2 – Triads

### 3.1 Introduction

Unlike most other probe chord studies, in this study we systematically explore the ways in which the frequency content of the context stimulus can affect the perceived dissonance/stability of triads. This section details the probe triad blocks of Experiments 1 and 2 in which the perceived stability and goodness-of-fit of major, minor, diminished and augmented triads given the context of a number of different scales is tested. The scale-tone tertian triads of the diatonic, harmonic minor and jazz minor scales are probed in the first experiment. We expect SPCS, triad type (to account for the consonance/dissonance of different triad types) and prevalence (represented by the frequency of occurrence of relative pitch classes within an appropriate – as explained in Section 3.2.2 below – tonal-harmonic corpus) to together predict our data, with effects such as pitch height also influencing ratings for individual trials. A second experiment includes two additional scales for which prevalence data are unavailable, namely: harmonic major and double harmonic, from which the major, minor, diminished and augmented scale-tone tertian triads are probed. We note that Chapter 2 tested probe tones of six scales in Experiment 2. We do not test probe triads in the four of these scales that are not *heptatonic* (that are not seven-note scales) as tertian triads are not well defined for non-heptatonic scales.

## 3.2 Overview of the Experiments and Models

### 3.2.1 Hypotheses

#### Experiment 1

Given Krumhansl's assumption that goodness-of-fit ratings directly measure musical stability, we should expect that our ratings of stability largely resemble those of goodness-of-fit. Pilot data suggested however that some particular triads, i.e., the leading-tone triads (triads rooted on the seventh degree of the scale), exhibit significantly lower stability than goodness-of-fit. This observation is not at odds with music-theoretic ideas, where the leading-tone of a scale leads strongly to the tonic, rendering it very unstable, but is still a member of the scale, so may fit the context reasonably well. Accordingly, our first hypothesis was that:

H1: Ratings of stability differ insignificantly from ratings of goodness-of-fit, apart from in a small number of cases that reflect music-theoretic ideas or tonal-harmonic musical practice.

We expect that SPCS and triad type will model the ratings well, but we assume that enculturation will also affect results. Accordingly, our second hypothesis reads:

H2: Perceived goodness-of-fit and perceived stability of probe triads may be modelled by the SPCS between the pitches of the context and the probe, and the consonance/dissonance of the probe and the statistical prevalence in Western music of the probe within the context scale.

Finally, considering that Krumhansl found that participants with less musical training responded more to pitch height cues, we expect that the musical sophistication of the participants will affect the degree to which they respond to the predictors in our model; that is,

H3: Significant interaction effects exist between musical sophistication and other predictors in such a model.

#### Experiment 2

Considering the scales used in experiment two are less familiar, we do not expect to see differences between goodness-of-fit and stability ratings on specific notes. Our first hypothesis is thus simplified:

H1: Ratings of stability differ insignificantly from ratings of goodness-of-fit.

Given that we cannot test for prevalence for these scales (and, due to the unfamiliarity of scales would not expect it to affect our results as strongly as in the first experiment even if we were able to test it), our second hypothesis reads:

H2: Perceived goodness-of-fit and perceived stability of probe triads may be modelled by the SPCS between the pitches of the context and the probe and the consonance/dissonance of the probe.

Finally, our third hypothesis remains from Experiment 1:

H3: Significant interaction effects exist between musical sophistication and other predictors in such a model.

### 3.2.2 Analysis

As for the probe tone data, the *R* package *brms* (Bürkner, 2017, 2018) was used to conduct Bayesian ordinal (cumulative logit) mixed effects regressions to test hypotheses 2 and 3, and to test for an observed tonal hierarchy. All continuous independent variables were *standardized* – centered at 0 and scaled to have a standard deviation of 1. The models were run using

$$\text{student\_t}(3, 0, 2.5)$$

(a *t*-distribution with 3 degrees of freedom, with mean of 0, scaled by 2.5) as a weakly informative prior. For the population-level effects that are *significant* in the models, along with their conditional effects and the intercepts, are displayed in a table. *brms's hypothesis test* is also run for all significant effects in the models to quantify their evidence ratios. To visually test for any systematic discrepancies between the observed data and the model predictions (Gelman et al., 2013) *brms's ppcheck* (a *posterior predictive check*) is used. A Bayesian version of the McKelvey-Zavoina pseudo- $R^2$  value (McKelvey & Zavoina, 1975) is also calculated.

For both experiments, we test initially for each scale whether or not a tonal hierarchy is observed for average ratings (whether fit or stability). This is achieved using model comparison via the leave-one-out cross validation information criterion (LOOIC), introduced in Section 2.2.3 in the previous chapter. We then test Hypothesis 1, concerning the similarity of fit and stability ratings. Following this is a descriptive model of the data concerning the accuracy with which SPCS is able to predict ratings, averaged over all trials, for each probe-context combination (this descriptive model is not used to test hypotheses). Hypotheses 2 and 3 are tested using a model of by-trial predictors for ratings including things like pitch height, recency

and trial number, as well as the key variables of interest – Prevalence and SPCS. For Experiment 1 this model is labelled *Model 3.1*, and for Experiment 2 it is labelled *Model 3.2*.

A model equivalent to Model 3.2 is run, but for the combined data sets for Experiments 1 and 2, and is labelled *Model 3.3*. Finally, exploratory analyses – using extensions of Models 3.1, 3.2, and 3.3 are included.

Significance values and effect sizes are interpreted as in the probe tone data.

The following subsections detail each of the above tests and models, as well as the latter's predictors.

### **Test for tonal hierarchies**

Before the hypotheses are tested we test for the emergence of a tonal hierarchy from ratings for each scale. To do this we run and compare two simple Bayesian ordinal mixed effects models. The models differ in their predictors of ratings, which, for the first model consist only of the probe, i.e., a concatenation of root with triad type, and for the second model only of the intercept (in both cases as both fixed and random effects). The two models are compared via cross-validation. If the model with probe significantly outperforms the model of the intercept (the null model) then this suggests that a tonal hierarchy was observed for the scale.

### **Comparison of fit and stability ratings (H1)**

A Mann-Whitney *U*-test was run to compare the fit to the stability ratings for each of the seven non-chromatic tertian triads for each of the three scales for Experiment 1, as well as six specific chromatic triads (detailed in Section 3.3.1 below), and for Experiment 2, five tertian triads from the double harmonic minor scale and seven from the harmonic major scale. The hypothesis is confirmed if, after Bonferroni corrections, the *p*-values resulting from the *U*-test are under .05 for triads for which the significantly greater perceived fit or stability aligns simply with music-theoretic discourse. Plots of bootstrapped fit and stability ratings for each non-chromatic (tertian) probe-triad combination are included in this section to accompany the *U*-test results. As mentioned above, H1 the individual scale-probe combinations for comparisons of fit to stability rather than overall differences in the predictors' effects between these two types of ratings, which are assessed using comparisons of alternative versions of the Bayesian mixed effects model.



### Descriptive model

Before hypotheses 2 and 3 are presented, SPCS is used to predict the average ratings for each scale-probe combination under the assumption that fit and stability do not differ significantly. In this way SPCS may be more directly compared to existing models of probe tone ratings in which average ratings of probes are modelled and it makes possible some useful summary visualisations. SPCS is introduced in more detail under the next heading.

### Testing H2 & 3 with Bayesian ordinal mixed effects models

Considering our first experiment as it was preregistered, following the recommendation of Barr et al. (2013) we ran the models with the maximal random effects structure driven by the design of the experiment, including random effects on participants with respect to Triad, SPCS, Prevalence, Primacy, Recency, Melodic Continuity, Count, and Relative Height; and their correlations.

Population-level effects considered in Model 3.2 are:

- *SPCS*: The spectral pitch class similarity of the aggregated pitches of the context and the aggregated pitches of the probe.
- *Triad*: The triad type of the probe, dummy coded with the major triad as the reference.
- *Recency*: Coded as 1 when one of the pitches of the probe matches the final pitch of the context, and 0 otherwise.
- *Primacy*: Coded as 1 when one of the pitches of the probe matches the initial pitch of the context, and 0 otherwise.
- *MelCont*: The melodic continuity of the probe from the context – the maximum number of consecutive intervals in a single direction that can be traced back from any of the three pitches of the probe (may take integer values from 1 to 7, given the one octave range).
- *Previous*: The rating given to the previous trial.
- *Count*: The sum of the number of occurrences in the experimental stimulus of the three pitches of the probe at the time of the trial.
- *RelHeight*: The sum of the pitch heights of the three pitches of the probe relative to its context (measured in semitones above the lowest pitch of the context. May take integer values from 6 to 24).
- *RelHeight<sup>2</sup>*: The relative pitch height of the probe squared.

- *Height*: The sum of the pitch heights of the pitches of the probe.
- *Height<sup>2</sup>*: The square of the sum of the pitch heights of the pitches of the probe.
- *TrialNo*: Trial number.
- *TrialNo<sup>2</sup>*: Trial number squared.
- *InContTrialNo*: Trial number within the group of trials of the same context scale.
- *Task*: Coded as  $-0.5$  if the participants rate fit and  $0.5$  if the participants rate stability.
- *BlockOrder*: Coded as  $-0.5$  if the probe tone block is presented first and  $0.5$  if the probe triad block is presented first.
- *MusSoph*: The musical sophistication of the participant, as measured by the Goldsmith Musical Sophistication Index.

*Height<sup>2</sup>* and *TrialNo<sup>2</sup>* are included in order to enable the modelling of a single bend in the distribution of the effect of *Height* and *TrialNo* (or of interactions with those effects) on ratings.

As a reminder, the SPCS of two pitch class sets is the cosine similarity between their spectral components after the application of Gaussian smoothing. SPCS's calculation is no different therefore whether one pitch class set consists of the spectra of a single tone, or of the three tones of a triad. A smoothing width of 10 cents (10% of a 12-TET semitone) was used. Smoothing widths of 6 and 14 cents were also tested but resulted in a marginally less well-fitting model; furthermore, 10 cents is close to values previously optimized in related experiments such as Milne, Laney, et al. (2015) and Milne et al. (2016). More information on the calculation of SPCS can be found in Appendix A. It is worth noting that removing the Gaussian smoothing, and more importantly, the spectral components of SPCS would reduce it to something akin to our *ScaleTone* effect, which simply counts the number of pitch classes shared by the two pitch class sets – the context scale, and the probe.

Model 3.1 also includes

- *Prevalence*: The frequency of occurrence of the probe in a corpus appropriate for the context scale.

The appropriate corpus for each context scale is not immediately clear. As discussed in Chapter 2, listening background questions included in the questionnaire completed by all participants suggest they listen to pop/rock music much more than

classical music (out of 63 participants, 26 listed rock and/or pop music compared to 7 classical, 9 both and 21 other). Accordingly, rather than using classical music corpora such as those used in Krumhansl (2001), we use Temperley's (2013) RS200 corpus. The RS200 corpus consists of the RS 5x20 corpus (De Clercq & Temperley, 2011) of the top 20 songs on the Rolling Stone magazine's list of the "500 Greatest Songs of All Time," from each decade from the 1950s through the 1990s (minus one, which was removed due to an absence of triadic harmony), with the addition of the next 101 songs from the list. Similarly to the probe tone block, our prevalence values are calculated from the frequency of occurrence of chords built on the probe triad in the songs that make up the major-like cluster for the diatonic scale, and the minor-like cluster for the harmonic-minor and jazz-minor scales, from Temperley and De Clercq's 2-cluster solution for the melodic corpus data. We obtained the statistical prevalences of the chords in the corpus using harmonic analyses available on Temperley's website (Temperley, n.d.) and classification of the songs into major-like or minor-like clusters, provided to us via personal communication (D. Temperley, personal communication, June 6, 2018).

As discussed in Chapter 2, in the preregistration (for Experiment 1. More details follow under Section 3.3 below), we intended to test all possible two-way interactions between all these variables. Such a model would be unfeasibly complex, both to computationally fit and to understand. Accordingly, we split the variables into two groups – those which represent a feature of the stimulus, and those which may affect the relative influence of such a feature on the participant's rating. Each variable of the second group interacts with each variable of the first group. The second group consists of MusSoph, TrialNo, TrialNo<sup>2</sup>, InContTrialNo, Task, and BlockOrder; the first group comprises the remaining effects.

To remind the reader, we hypothesize (H2) that perceived goodness-of-fit and perceived stability of probe tones may be modelled by the SPCS of the aggregated pitches of the context and the probe, and (for Experiment 1 only) the statistical prevalence in Western music of the probe within the context scale. To answer H2, an associated reduced model is run, and the full model and the reduced model are compared via cross-validation. In Experiment 1, the reduced model differs from the full model by the absence of Triad, SPCS and Prevalence, and for Experiment 2 only by the absence of Triad and SPCS (since the models of Experiment 2 do not include Prevalence). The hypothesis is confirmed in each case if the reduced model is significantly outperformed. In contrast, H3 concerns the significance of the interaction

of several effects with musical sophistication in the model. Accordingly, rather than by model comparison, H3 is confirmed if the 95% credibility interval lies entirely above or below zero for one or more interaction effects with musical sophistication.

### Exploratory analyses

An exploratory analysis considers whether adding an effect of the Inversion of the triad improves the models. The triads are dummy coded with root position as the reference, (and first and second inversion the other options). As well as Major, Minor and Diminished triads, for which these inversions are easily understood, the same is done for Augmented triads, for which they are not as clearly associated. Using the C Augmented triad for example – C-E-G $\sharp$  – ‘C’ is the root, ‘E’ the third, and ‘G $\sharp$ ’ the fifth. We consider the triad E-G $\sharp$ -C to be in first inversion and the triad G $\sharp$ -C-E to be in second inversion, despite the fact that in 12-TET, all three triads, out of context, are equivalent, given that in all of them all notes are separated by 4 semitones (a major third). Given that our scales were defined in terms of harmonic tonality, however, these triads are not the same, as reflected in the spelling of the pitches in the triad.

In order to explore whether SPCS accounts for any differences in ratings more complex than the number of pitch classes in the probe triad that are also in the context scale, an additional exploratory analysis considers an additional effect of *Scale-Tone* (whether or not the probe tone is heard in the context melody). This consideration is explored by adjusting Models 1, 2 and 3 by replacing SPCS with ScaleTone and comparing via LOOIC to Models 1, 2 and 3 respectively. A model equivalent to Model 3.1, but without Prevalence is also compared with a model equivalent to it but with ScaleTone instead of SPCS.

## 3.3 Experiment 1: Diatonic, Harmonic Minor, Jazz Minor

The first experiment tested the perceived goodness-of-fit and stability of triads given the randomly ordered, uniformly distributed sounding of the pitches (sounded three times each) of three different context scales common to Western tonal music – the diatonic, harmonic minor and jazz minor – sounded with harmonic complex tones.

If we find that the results cannot be modelled accurately by SPCS, this psychoacoustic description of the cognition of tonality is not supported. Participants, including both musicians and non-musicians, also completed the Gold-MSI Questionnaire in order to account for effects of musical sophistication upon their responses.

The experiment was pre-registered through Open Science, available at <https://osf.io/az6x8>. The preregistered experiment differs slightly, as we realised that the analysis could be potentially improved by the inclusion of a small number of effects not included in the pre-registration – namely,  $\text{RelHeight}^2$ ,  $\text{Height}$ ,  $\text{Height}^2$  and  $\text{Previous}$ . We believe that considering these added effects allows for a fuller description of the data. The analysis of the pre-registered model is shown only in Appendix C in Table C.1, along with the results of the hypothesis tests performed on that model, which are the same as for Model 3.1.

The preregistered model also includes an effect of Task – whether or not participants rated fit or stability – which also interacts with the same effects as SPCS, etc. An analysis of the preregistered model (via LOOIC comparison) revealed that the removal of this effect does not reduce the performance of the model. Model 3.1, without this effect then can predict ratings of both fit and stability, and is considerably simpler.

### 3.3.1 Method

The method differs from the probe tone block only by the probes. All tertian triads from within the context scale are probed, as well as 6 additional triads, comprising

1. one randomly selected major, minor and diminished triad that includes one or more pitches outside of the scale
2. the *tonic parallel* of each scale, i.e., for the diatonic scale, the minor triad (instead of the major) rooted on the scale's theoretical tonic – the RPC numbered '0' – and for the minor scales (harmonic and jazz), the major triad rooted on the theoretical tonic of each scale

all in a random inversion.

In a seven-note scale that spans one octave, three tertian triads occur in root position, two in first inversion, and two in second inversion. The identity of the triads of each inversion type changes when the RPC of the lowest pitch changes.

In order that no more root position triads are heard than first or second inversion triads, for each context scale, all scale-tone tertian triads are probed apart from one: the triad rooted on the lowest, second lowest or third lowest pitches. The row that is not probed is randomly determined and is consistent across context scale. This ensures that, if equivalence across inversion is assumed, each combination of probe triad and context is heard twice. For example, for pitch classes C, D, E, F, G, A and B one rotation has D as the lowest note so the triads on D, E and F occur in root position, those of G and A in second inversion and those on B and C in first inversion. When F is the lowest note, the triads of F, G and A occur in root position, B and C in second inversion and D and E in first inversion. Either the lowest, second lowest or third lowest triads are removed for both these rotations – i.e. either D and F, E and G or F and A respectively.

The order of probes is randomized within the contexts, whose order is also randomized. For the probe triads block, three randomly determined extra (chromatic – detailed above) probes are included. These may differ between participants. An additional set of specific chromatic probe triads (detailed above), representing chromatic triads that are commonly used in tonal-harmonic music plays at the end of the probe triad section. Within each trial, the order of notes in the context is also randomized. Randomly situated amongst the 7 probe triads in 6 of 7 modes (rotations) = 42 trials for each of the three context scales, 1 chromatic triad is probed on two occasions, leading to 44 trials. Concluding the block, 3 additional triads are probed twice each, leading to 50 trials for each of three context scales and 150 trials for the block. Before the experimental trials, each block begins with 6 practice trials, leading to  $72 + 150 + 6 \times 2 = 234$  trials for the whole experiment.

To remind the reader, between the two experimental blocks (a block of 72 probe tone trials and a block of 150 probe triad trials) the participants also completed a survey including the Goldsmith MSI Questionnaire, in order to obtain an index for musical sophistication to be used as a variable in analysis. Additional demographic questions followed the Goldsmith MSI Questionnaire to facilitate future analysis of possible effects of enculturation (these are not analysed here).

### 3.3.2 Results

For all results, triads of different inversion are not considered to be different triads; thereby simplifying the models and their interpretation. A justification could

be found after consideration of the inversional equivalence suggested in Mathews et al. (1988), and in Roberts and Shaw (1984). Where in a LOOIC comparison between models equivalent but for the presence or absence of inversional equivalence (discussed in Section 3.3.2) the model with inversional equivalence performs significantly better, the justification is strengthened. In the following analyses “triad” refers to an equivalence class of triads consisting of the same pitch classes in any inversion. We reiterate here also that triads are defined not by the pitch heights of their pitches, or just of their root, but by their triad type (major, minor, diminished or augmented) and the RPC of their root within the context scale they follow.

### Test for tonal hierarchies

Comparisons of Bayesian models of average ratings suggest that a tonal hierarchy is observed for all three scales. For each scale a null model of only the intercept is compared to an alternative model of an interaction between triad type and the RPC of the triad’s root. In each case the alternative model performed significantly better than the null, as shown in Table 3.1.

TABLE 3.1: LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 1

Scale	Null – Alternative LOOIC	SE	Signif
Diatonic	826.6	63.8	Yes
Harmonic minor	859.4	66.6	Yes
Jazz minor	661.8	56.6	Yes

### Comparison of fit and stability ratings

To give an overall picture of the difference between fit and stability the average ratings of fit and stability for all participants are shown in Fig. 3.1 (diatonic scale), Fig. 3.2 (harmonic minor scale), and Fig. 3.3 (jazz minor scale). Error bars are from 95% confidence intervals obtained from 1000 bootstrapped samples.

These diagrams suggest that

- The leading-tone triads of the harmonic minor and jazz minor scales are given higher fit than stability ratings.
- Though the minor tonic and the major dominant of the jazz minor scale are the equally best fitting triads, the major triads, the dominant and subdominant, are the most stable

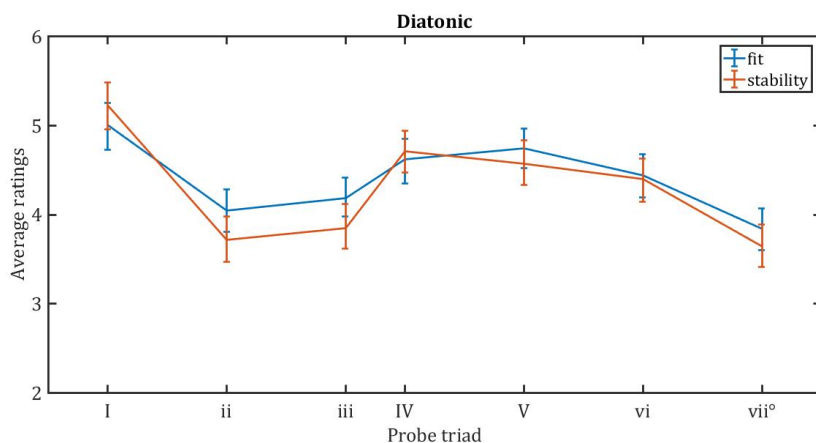


FIGURE 3.1: Average fit vs stability ratings for probe triads after the diatonic context. Triads are labelled by the Roman numeral for the scale-tone of their root.

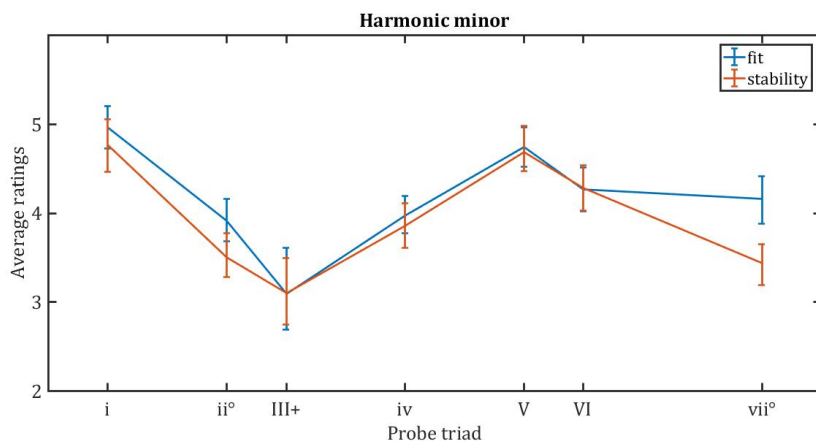


FIGURE 3.2: Average fit vs stability ratings for probe triads after the harmonic minor context

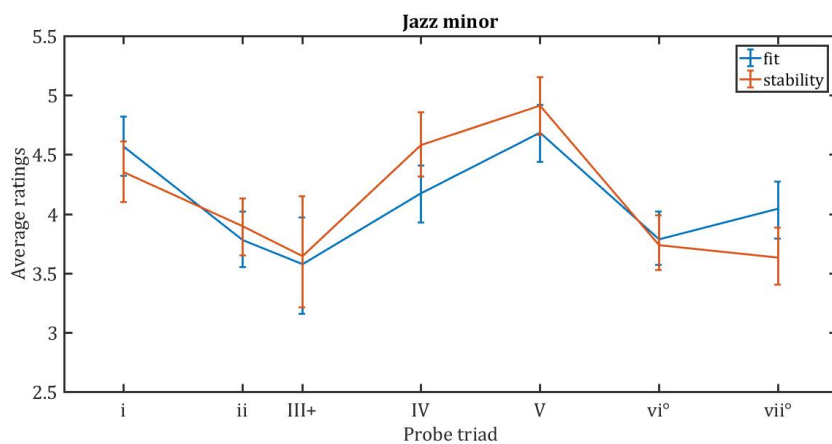


FIGURE 3.3: Average fit vs stability ratings for probe triads after the jazz minor context



For the scale-tone tertian triads and the tonic parallels (triads rooted on the scale's theoretical tonic RPC 0, with a third of the "other" size – major or minor) for the three context scales (averaged over inversion), we conducted a Mann-Whitney *U*-test comparing the *Z*-scores of ratings of fit to ratings of stability. This makes up  $3 \times (7 + 1) = 24$  context-probe combinations. After applying Bonferroni corrections, we found that for the jazz minor scale both the tonic and leading-tone triads received significantly higher fit than stability ratings.

Though significant differences are found for particular probe triads within one of the scales, we cannot so far say whether or not our Bayesian ordinal regression model benefits from the inclusion of an effect of task – whether ratings were of fit or stability. This question is explored in section 3.3.2. It is found that the model without Task does not perform significantly worse. Accordingly, in the following model we average fit and stability ratings together as "average ratings".

### **Descriptive model**

Before the ordinal mixed effects models used to test the second and third hypotheses, a linear model was run comparing observed ratings for each scale-probe combination, averaged over all other variables, to a model using SPCS and Triad.<sup>1</sup> Figures 3.4, 3.5 and 3.6 plot the SPCS prediction from this model against the average observed ratings for the diatonic, harmonic minor and jazz minor scales respectively. Chromatic triads (triads including non-scale-tones) were included in the model but are not shown in Figures 3.4–3.6 in order to keep the plots simple and considering that nothing of note was found of these triads in any analyses.

---

<sup>1</sup>Given the ordinal nature of the dependent variable, a linear model is not ideal; however it makes for simpler interpretation for these data. This, and the fact that many previous probe-tone experiments used linear models which we wish to compare to, is why we use a linear model. Inspection of a plot of the residuals suggested that a linear model is appropriate for our data.

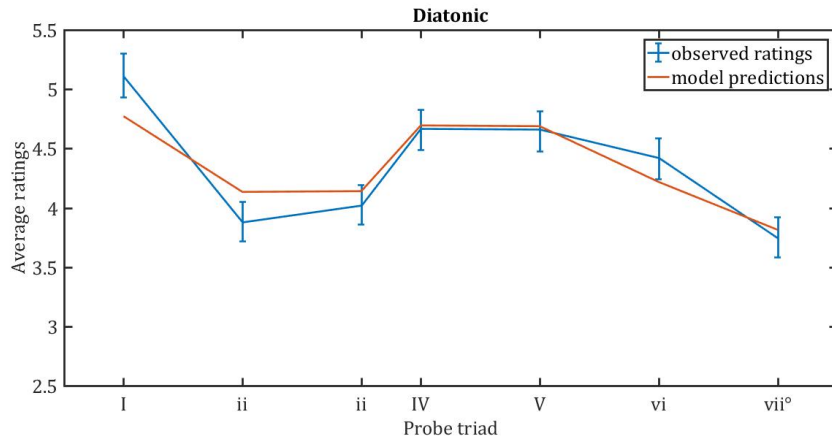


FIGURE 3.4: Average ratings for probe triads after the diatonic context compared to predictions due to SPCS and Triad

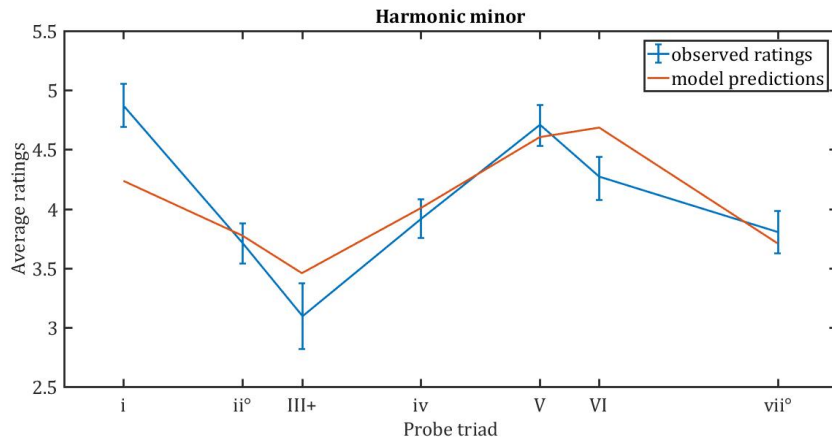


FIGURE 3.5: Average ratings for probe triads after the harmonic minor context compared to predictions due to SPCS and Triad

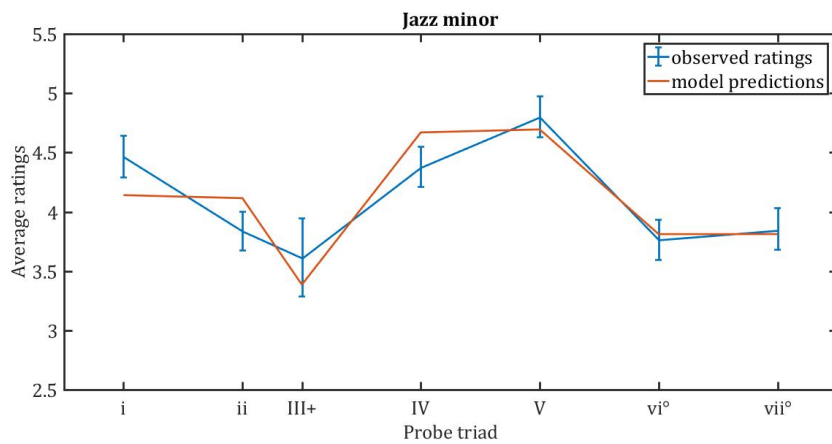


FIGURE 3.6: Average ratings for probe triads after the jazz minor context compared to predictions due to SPCS and Triad

Observed ratings for (most common) theoretical tonics for all three scales were clearly higher than predicted by a model of SPCS and triad type. For the diatonic scale the major tonic triad was rated higher than the other triads. For the harmonic minor scale 'i' and 'V' were rated higher than the other triads, and 'III+', the only augmented triad, was rated lower than the other triads. For the jazz minor scale, 'i', 'V' and 'IV' were rated higher than the other triads. Only the diatonic scale has a single clear tonic (one triad rated higher than the others, without any error bars overlapping), and for the jazz scale 'i' is rated lower than 'V', though the error bars overlap. Note that the error bars are wider for the augmented triads. This is because inversions of augmented triads are all equivalent.<sup>2</sup> Each inversion of each triad type was sounded the same amount of times. Since all other triads come in three different inversions, augmented triads were heard one third as many times as the other probe triads.

Although major triads are often rated higher than minor, and minor higher than diminished and augmented, this is not always the case. We can see that this added complexity in ratings is only partially predicted by SPCS.

The resulting model fits the data less well than our model of only SPCS for our probe tone data for the same set of scales (see Chapter 2):  $R^2 = .73$ , adjusted .72, in comparison to .85, adjusted .84.

In contrast to these simple linear models, the ordinal mixed-effects models – detailed in the next section – model ratings given for each individual trial.

### **H2&3: Model 3.1**

Two alternative models were run, differing only in the presence or absence of an effect of Task (fit or stability). A LOOIC comparison of the two models favours the model without Task (Table 3.4) and it is chosen as Model 3.1 accordingly. For brevity only the significant effects are shown for this model in Table 3.2, and in the tables of later ordinal models, along with any associated conditional main effects and the intercepts. A table including all effects is included in Appendix C (Table C.2). For the categorical effect of Triad, 'Min', 'Aug' and 'Dim' label the effect of the triad having the identity minor, diminished or augmented, as opposed to major.

---

<sup>2</sup>Augmented triads are built from two major thirds of 4 degrees of 12-TET. The remaining interval to the octave is a diminished fourth, also of 4 degrees. Thus the augmented triad, in any inversion, comprises adjacent intervals of 4 degrees.

TABLE 3.2: Model 3.1 significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.24	0.14	-3.51	-2.97	5072	1.00
Intercept[2]	-1.77	0.13	-2.03	-1.51	5053	1.00
Intercept[3]	-0.54	0.13	-0.80	-0.29	5003	1.00
Intercept[4]	0.12	0.13	-0.13	0.38	5007	1.00
Intercept[5]	1.33	0.13	1.08	1.59	5107	1.00
Intercept[6]	2.89	0.14	2.62	3.16	5311	1.00
MusSoph	-0.11	0.11	-0.33	0.10	3777	1.00
RelHeight	-0.11	0.25	-0.60	0.38	5393	1.00
Previous	0.24	0.05	0.14	0.33	6619	1.00
Recency	0.24	0.09	0.08	0.41	7526	1.00
SPCS	0.39	0.07	0.26	0.53	4852	1.00
Count	0.03	0.06	-0.09	0.15	9259	1.00
Minor	-0.43	0.13	-0.69	-0.18	4815	1.00
Diminished	-0.59	0.17	-0.92	-0.25	5063	1.00
Augmented	-1.06	0.21	-1.47	-0.65	5611	1.00
Prevalence	0.32	0.07	0.18	0.46	5939	1.00
TrialNo	0.09	0.07	-0.05	0.23	6101	1.00
InContTrialNo	-0.08	0.06	-0.20	0.04	5485	1.00
Order	-0.04	0.23	-0.50	0.42	3272	1.00
MusSoph:SPCS	0.22	0.06	0.10	0.35	5311	1.00
MusSoph:Minor	0.26	0.11	0.03	0.48	4749	1.00
MusSoph:Diminished	0.52	0.15	0.22	0.82	5122	1.00
MusSoph:Augmented	0.14	0.18	-0.22	0.50	5313	1.00
MusSoph:Prevalence	0.24	0.06	0.12	0.37	6540	1.00
RelHeight:TrialNo	-0.47	0.24	-0.95	-0.02	4723	1.00
Count:InContTrialNo	-0.08	0.03	-0.14	-0.03	11670	1.00
Count:Order	-0.30	0.10	-0.49	-0.11	9238	1.00

Significant population level effects for Model 3.1 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

Intercepts represent 'cutpoints' between ordinal values of ratings and the Intercept values are of a latent continuous variable; the Estimate value corresponds to the change in this latent variable that is associated with an increase of 1 standard deviation in the value of the effect for continuous variables, or with an increase from a value of 0 to a value of 1 for binary variables. We interpret effects with a magnitude of about 0.2 to be small, those of about 0.5 to be medium, those of about 0.8 to be

large.

As for the probe tone models, intercepts represent “cutpoints” between ordinal values of ratings and the Intercept values are of a latent continuous variable; the Estimate value corresponds to the change in this latent variable that is associated with an increase of 1 standard deviation in the value of the effect. For example, given the Estimate value of  $-1.06$  for Augmented (the estimated effect of a triad being Augmented rather than Major), the difference between a major chord and a diminished chord would take a rating of 4 (an estimate between  $-0.54$  and  $0.12$ ), for example, to a rating of 3 (estimate between  $-1.77$  and  $-.54$ ). The size of Estimates may be interpreted as thus:  $0.8$  – large,  $0.5$  – medium,  $0.2$  – small (values are first divided by  $1.6$  (Amemiya, 1981) – the value of the variance of a Gaussian distribution that best approximates a logistic distribution with a variance of  $1$  – in order to approximate the SD of the latent variable. This results in Pearson’s  $r$  coefficient, wherein  $0.5$  – large,  $0.3$  – medium,  $0.1$  – small. The values given above approximate these values multiplied by  $1.6$ ).

SPCS and Prevalence are both significant in the positive direction as main effects and as interactions with MusSoph (musical sophistication), suggesting that musician participants respond more strongly both to SPCS and to Prevalence. Count is significant also as an interaction with InContTrialNo (in-context trial number) and BlockOrder in the negative direction. This suggests that when the probe triad block is presented after the probe tone block, and for earlier trials within a single context scale, the sum of the number of occurrences in the experimental stimulus of the three pitches of the probe at the time of the trial has a stronger effect on ratings. Additionally, a negative interaction between RelHeight and TrialNo is also significant, suggesting that the effect of relative pitch height weakens as the experiment progresses.

Finally, from interactions between MusSoph and Triad, we can deduce that more musically sophisticated participants rate major triads significantly higher than minor and diminished triads. The latter confirms our third hypothesis. To test for differences in ratings between other triads, interacting with MusSoph, we run hypothesis tests in *brms*. We confirm that  $\text{MusSoph:Dim} > \text{MusSoph:Min}$ , with an evidence ratio of  $28.78$ , which constitutes strong evidence and that  $\text{MusSoph:Dim} > \text{MusSoph:Aug}$ , with an evidence ratio of  $41.7$ , which constitutes very strong evidence. A plot (Figure 3.7) of the conditional effect of MusSoph:Triad details this interaction.

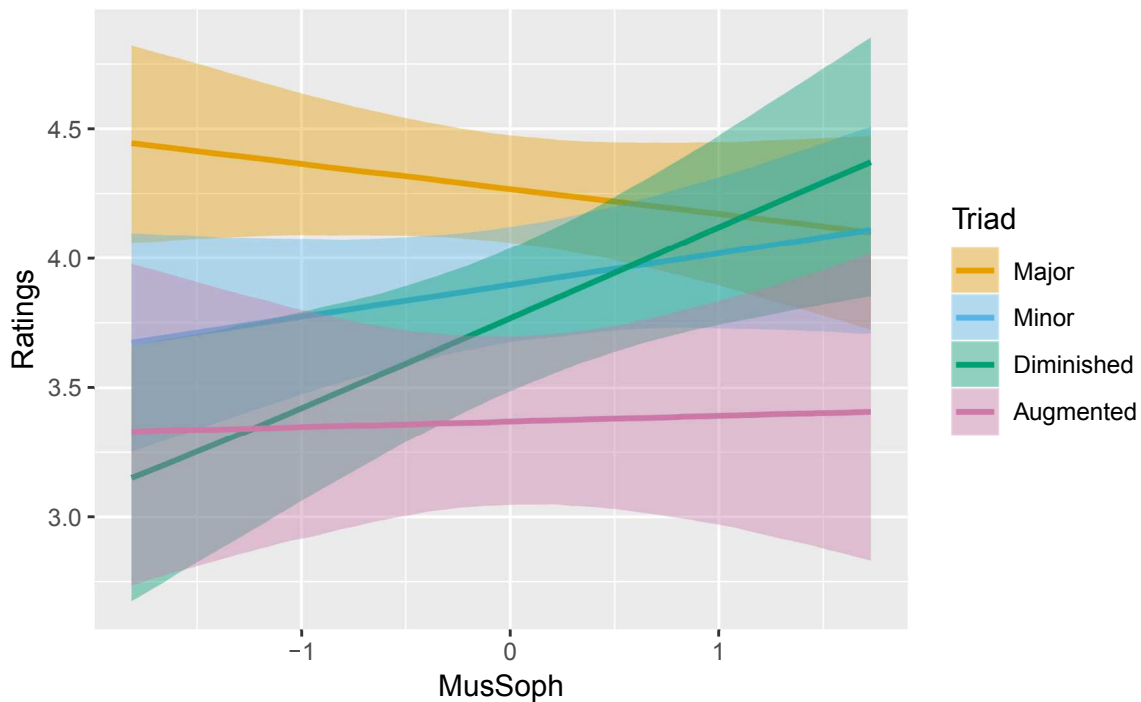


FIGURE 3.7: Conditional effect of MusSoph:Triad for Model 3.1

Table 3.3 comprises the results of hypothesis tests on all significant effects, many of which, involving comparisons between non-major triads, were not displayed in Table 3.2.

TABLE 3.3: Evidence ratios for all significant effects of Model 3.1

		evidence ratio	strength
Major	> Minor	1999	very strong
Major	> Dim	11999	very strong
Major	> Aug	> 11999	very strong
Minor	> Aug	1332.33	very strong
Dim	> Aug	92.75	very strong
SPCS	> 0	> 11999	very strong
Recency	> 0	399	very strong
Prevalence	> 0	> 11999	very strong
Previous	> 0	> 11999	very strong
MusSoph:Min	> MusSoph:Maj	72.62	very strong
MusSoph:Dim	> MusSoph:Maj	1713.29	very strong
MusSoph:Dim	> MusSoph:Min	28.78	strong
MusSoph:Dim	> MusSoph:Aug	41.7	very strong
MusSoph:SPCS	> 0	3999	very strong
MusSoph:Prevalence	> 0	> 11999	very strong
RelHeight: TrialNo	< 0	47.58	very strong
Count: InContTrialNo	< 0	1713.29	very strong
Count: BlockOrder	< 0	704.88	very strong

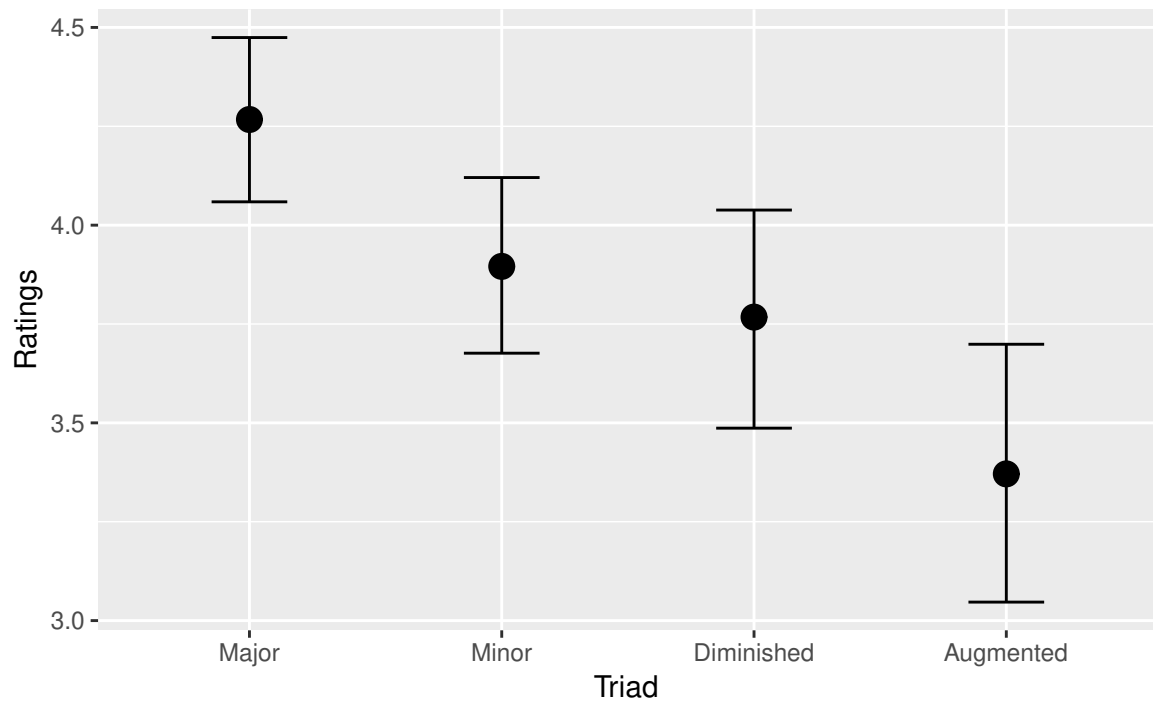


FIGURE 3.8: Conditional effect of Triad for Model 3.1

A plot (Figure 3.8) of the conditional main effect of triad reveals the expected consonance ordering for the triad types, though Figure 3.7 suggests that this ordering is not demonstrated in participants with higher musical sophistication.

The same model was run without SPCS, Prevalence and Triad as predictors and judged by LOOIC performed significantly worse (see Table 3.4 above: row 2). Our hypothesis is thus confirmed: perceived goodness-of-fit and perceived stability of probe triads may be modelled by the SPCS between the pitches of the context and the probe, the consonance/dissonance of the probe and the statistical prevalence in Western music of the probe within the context scale.

TABLE 3.4: LOOIC comparisons for Experiment 1 Bayesian ordinal mixed effects models

Model compared to Model 3.1	Model – Model 3.1	LOOIC	SE	Signif
+ Task		0.8	5.8	no
– Triad – SPCS – Prevalence		2266.4	104.8	yes
– SPCS		401.4	52.6	yes
– Prev		242.8	41.6	yes
– Triad		543.5	53.8	yes
+ Inversion		–25.0	20.8	no
– SPCS + ScaleTone		9.0	10.2	no
Model compared to Model 3.1 – Prevalence	Model – (Model 3.1 – Prevalence)	LOOIC	SE	Signif
Model 3.1 – Prevalence – SPCS + ScaleTone		90.2	14.4	yes
Model 3.1 – SPCS		166.8	69.4	yes

H2&3 were similarly confirmed for the preregistered model, as shown in Appendix C.

The ppcheck plot for this model confirms the validity of the model. With a pseudo- $R^2$  value of .45, this model does not fit the data as well as the model for the probe tones for this experiment (discussed in Chapter 2, which had a pseudo- $R^2$  of .57).

### Exploratory analysis

The effects of SPCS and Prevalence are intended to account for two distinct processes that account for the cognition of harmonic tonality. Prevalence models a top down process based on long term statistical learning of the frequency of occurrence of RPCs in tonal-harmonic music whereas SPCS is a bottom-up process based on a psychoacoustic response to the frequency content of the stimulus. SPCS may also influence familiarity because the statistics of music may themselves be shaped by psychoacoustic features such as SPCS. We explore the relationship between SPCS and Prevalence by running two models, equivalent to Model 3.1, but with Prevalence removed, or with SPCS removed, respectively. With Prevalence removed, SPCS has an effect size of 0.5, which is larger than the effect sizes of SPCS or Prevalence in Model 3.1, but smaller than their sum (the model is shown in Appendix C in Table C.3). With SPCS removed, Prevalence has an effect size of .46, similar to that of SPCS in the most recently mentioned model. Both models are significantly outperformed by Model 3.1. (Table 3.4, rows 3 and 4). This suggests that Prevalence and SPCS account for different aspects of responses. Finally, a model equivalent



to Model 3.1 but with Triad removed is run, which also underperforms Model 3.1 (Table 3.4, row 5).

We might ask then what information SPCS can provide beyond simply a prediction based on the number of pitches in the probe triad that were also heard in the context. From our descriptive model we know that there is more complexity in ratings than can be explained by this binary partition: in each scale, one RPC (the tonic) is rated significantly higher than the others. We can see that changing a single RPC in the context from the diatonic to the harmonic minor scales, for example, influences not only the ratings of RPCs that shifted from being scale-tones to chromatic RPCs and vice versa, but also of the tonics' RPCs. Though the tonics are not predicted by SPCS, one can see that within the non-tonic scale-tones and within the non-scale-tones there are differences in SPCS that do appear to reflect differences in ratings.

We define the additional predictor *ScaleTone*, coded as the number of pitch classes of the probe triad that are scale-tones, i.e., that are included in the context stimulus. We alter Model 3.1 by replacing SPCS with ScaleTone. In this altered model ScaleTone and the interaction between ScaleTone and MusSoph are both significant in the positive direction, and a LOOIC comparison between the models suggests no difference in out-of-sample predictive performance between the Model 3.1 and this alternate model (details shown above in Table 3.4, row 7). In a model with ScaleTone and without SPCS or Prevalence, ScaleTone and MusSoph:ScaleTone are again both positive and significant; however, this model is found to be significantly worse (shown in Table 3.4, one row from the bottom) than a model equivalent to this but with SPCS instead of ScaleTone. The significant population level effects and associated conditional effects for each of these models are included in Appendix C (Tables C.4 and C.5). We can conclude that SPCS predicts beyond the capacity of ScaleTone only when Prevalence is not also included as an effect.

A model comparison was also run to ascertain whether or not inversional equivalence should be assumed in our model. To compare with Model 3.1 an additional model was run that differed only by the inclusion of 'Inversion' as an effect in addition to 'Triad'. We find that the assumption of inversional equivalence does not significantly change the performance of the model (see Table 3.4). Since it simplifies the model we apply inversional equivalence to our models.

We also considered more complex models of Recency and Primacy, namely the pitch distance of the probe in semitones of from the final and initial pitches of the

context melody, respectively, and the square of these values. We found that these did not improve the model as judged by LOOIC.

### 3.3.3 Discussion

Considering our exploratory analyses, when SPCS replaced ScaleTone in a model including Prevalence the performance of the models is judged via LOOIC to be insignificantly different, but if the models do not include Prevalence the model with SPCS is judged to be significantly better. This suggests that SPCS is a better measure than ScaleTone of a mechanism at least partially accounted for by Prevalence (which we can safely assume is related to long-term statistical learning) as well as by another mechanism that is not predicted by Prevalence. Where Model 3.1 significantly outperformed models equivalent but for the removal of Prevalence and SPCS respectively we understand that Prevalence and SPCS are each able to account for a mechanism the other is not.

We consider now the results of the Mann-Whitney  $U$ -test for H1. Although the leading-tone triad of the jazz minor does not include any pitches not heard in the context, since it is a diminished triad, and also is rooted on the leading-tone – the least stable RPC, we are not surprised to see that it received significantly higher fit than stability ratings. We cannot explain however why the leading-tone triads (also diminished) of the diatonic and harmonic minor scales did not also receive significantly higher fit than stability ratings. Considering the minor triad of the jazz minor scale, we note that as the tonic of the scale we expect high stability – as predicted by SPCS, whereas as a minor triad, we expect moderate stability. Where the triad was given significantly higher fit than stability ratings, we might wonder if stability is perhaps more affected by Triad than SPCS, however, no significant differences between fit and stability were found in a Bayesian mixed effects regression model.

Our confirmatory analysis suggests that SPCS, Prevalence, and Triad are all important in modelling tonal fit for triads, and our exploratory analysis suggests that a triad's inversion is not.

In the Section 3.2.2 we discussed the difficulty in appropriately using prevalence for our experimental design. While possible for these major and minor scales, this will not be possible for less common scales for which we can assume possible effects of familiarity will be substantially diminished. If there are clearly hierarchies of perceived fit or stability, and we find SPCS to contribute to a model for goodness-of-fit

or stability ratings for tones and after the context of less common scales, then this strengthens the argument for the utility of a psychoacoustic description of the cognition of harmonic tonality. Experiment 2 involves such a set of scales as stimulus.

### 3.4 Experiment 2: Less familiar scales

In Experiment 1, the common usage of the scales used for context meant that the effect of statistical learning needed to be considered in relation to the perceived fit/stability of probes sounded after the context scales. A second experiment was devised using as context less familiar scales for which this effect should be reduced, namely the harmonic major and double harmonic scales, which, notated in common modes beginning on C for convenience, are:

- Harmonic major: C D E F G A $\flat$  B
- Double Harmonic: C D $\flat$  E F G A $\flat$  B

In RPCs these correspond to:

- Harmonic major: 0 2 4 5 7 8 11
- Double Harmonic: 0 1 4 5 7 8 11

Both scales were tuned to 12-tone equal temperament.

If the results can be modelled accurately by SPCS, a psychoacoustic description of the cognition of tonality is supported. Participants, including both musicians and non-musicians, also completed the Goldsmith MSI Questionnaire in order to account for effects of musical sophistication upon their responses.

#### 3.4.1 Method

The stimulus differs from Experiment 1 not only by the context scales, but also by the triads probed. In the double harmonic scale there are two tertian triads that are not major, minor, diminished or augmented. These are not probed. For the harmonic major scale one row of triads is not probed as in the scales of Experiment 1. Chromatic triads probed for each scale comprise one major, minor and diminished triad containing at least one chromatic pitch class.

The experimental procedure differs in the number of trials in each block: The probe tone block includes  $12 \times 6 \times 2 = 144$  trials (analysed in Chapter 2); the probe

triad block includes  $(6 + 5) \times 7 + 3 \times 2 \times 2 = 89$  trials. Before the experimental trials, each block begins with 6 practice trials, leading to  $144 + 89 + 6 \times 2 = 245$  trials for the whole experiment.

## Participants

Thirty-two musicians (participants who reported having received 5 or more years of musical training) and 32 non-musicians were recruited for the experiment. Non-musician participants were first-year university students recruited through Western Sydney University School of Psychology and Social Science’s SONA system and received credit points towards their degrees for their participation. Musician participants were recruited via personal connection and received a \$30 reimbursement for their time and travel to the university campus. All participants reported normal hearing capabilities. Survey data were received for only 28 musicians and 29 non-musicians. Of these, four (musician) participants reported having received 10 or more years of musical training and one reported having absolute pitch. Participants had a mean age of 21.5 years, with a SD of 4.0 years. Of the 28 musicians 16 were female, and out of the 29 non-musicians 24 were female. This research was approved by the Western Sydney University Human Research Ethics Committee under the number H11908.

## 3.4.2 Results

### Test for tonal hierarchies

For both scales, a model of triad type interacting with triad root RPC performed significantly better than a model of only the intercept, as shown in Table 3.5. This suggests that tonal hierarchies were observed for both scales.

TABLE 3.5: LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 2

Scale	Null – Alternative LOOIC	SE	Signif
Harmonic major	318.2	50.6	Yes
Double harmonic	140.0	35.0	Yes

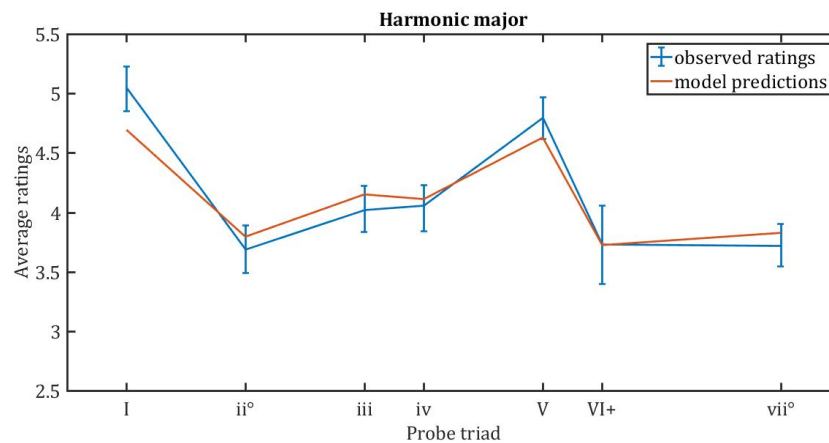


FIGURE 3.9: Average ratings for probe tones after the harmonic major context compared to predictions due to SPCS and Triad

### Comparison of fit and stability

A Mann-Whitney  $U$ -test was run comparing ratings of fit to ratings of stability for all tertian major, minor, diminished or augmented triads of both scales (numbering 7 for the harmonic minor, and 5 for the double harmonic). After Bonferroni corrections, no significant differences were found and our first hypothesis was thus supported. In the descriptive model below, fit and stability ratings are averaged together as “average ratings”, and task is not included as an effect in Model 3.2, which is used to test our second and third hypotheses.

### Descriptive model

A linear model was run with SPCS and Triad as predictors of ratings averaged over participants, task and trials.<sup>3</sup> Figures 3.9 and 3.10 show SPCS’s predictions for the six scales of Experiment 2, against the observed ratings, with 95% confidence intervals from 1000 bootstrapped samples.

For the harmonic major scale, we can see that the theoretical tonic (‘I’) and dominant (‘V’) triads are rated higher than the others. For the double harmonic scale triads ‘I’, ‘II’ and ‘iv’ are rated higher than triads ‘iii’ and ‘VI+’. Of all scales from Experiments 1 and 2 the harmonic major scale is most closely predicted by the linear model, which is unable only to predict the high ratings of triad ‘I’. As in Experiment 1, the augmented triads, ‘VI+’ on both scales, were played one third as many

<sup>3</sup>As in Experiment 1, inspection of a plot of the residuals suggested that a linear model is appropriate for the data averaged over participants.

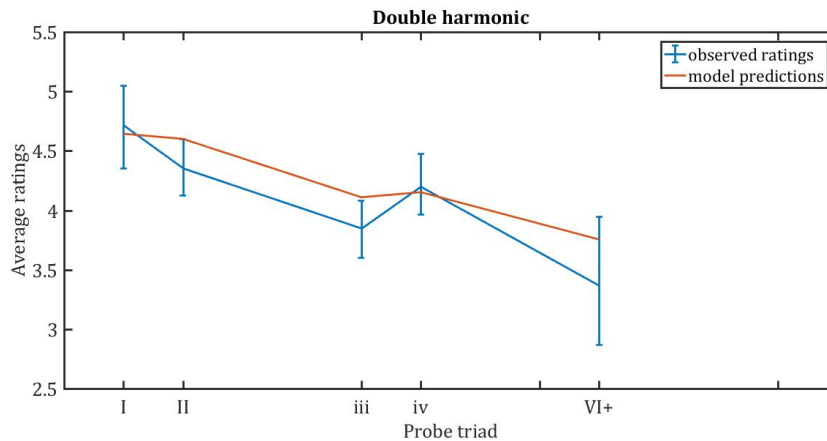


FIGURE 3.10: Average ratings for probe tones after the double harmonic context compared to predictions due to SPCS and Triad

times to participants as the other triads for the scale-tone triads.<sup>4</sup> The plots also suggest that the tonics of the exemplar modes of the scales function perceptually as tonics, having received slightly higher ratings than all other RPCs, though the error bars associated with the tonics overlap those of at least one other RPC.

For the two scales of Experiment 2, our model of SPCS and triad type is not as strong as the respective model for the scales of Experiment 1, with  $R^2 = .66$ , adjusted  $.64$  (though the model seems to fit the data for the harmonic major scale quite well).

### H2&3: Model 3.2

The complete model from Experiment 1 without Prevalence is run for the unfamiliar scales in Experiment 2. Table 3.6 details the effects significant in the model, along with the intercepts and associated conditional effects. A table including all effects is shown in Appendix C (Table C.6).

<sup>4</sup>Due to an undiagnosed error in the MATLAB script, slightly fewer augmented triads than that were heard by participants for the double harmonic scale.

TABLE 3.6: Model 3.2 significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.53	0.19	-3.90	-3.16	5837	1.00
Intercept[2]	-2.34	0.18	-2.70	-1.98	5919	1.00
Intercept[3]	-1.16	0.18	-1.51	-0.80	5965	1.00
Intercept[4]	-0.31	0.18	-0.66	0.05	5958	1.00
Intercept[5]	0.94	0.18	0.59	1.30	6049	1.00
Intercept[6]	2.46	0.19	2.10	2.83	6293	1.00
MusSoph	0.38	0.15	0.08	0.68	4300	1.00
Previous	0.32	0.07	0.18	0.46	8016	1.00
SPCS	0.27	0.06	0.15	0.38	10689	1.00
Minor	-0.72	0.14	-0.99	-0.46	8748	1.00
Diminished	-1.18	0.17	-1.51	-0.86	8011	1.00
Augmented	-1.26	0.20	-1.66	-0.86	8835	1.00
BlockOrder	-0.46	0.32	-1.10	0.16	5040	1.00
MusSoph:SPCS	0.15	0.05	0.05	0.24	9910	1.00
MusSoph:Minor	-0.38	0.12	-0.60	-0.15	8309	1.00
MusSoph:Diminished	-0.33	0.15	-0.63	-0.03	6917	1.00
MusSoph:Augmented	-0.53	0.16	-0.85	-0.21	9291	1.00
Minor:BlockOrder	0.46	0.22	0.03	0.90	8552	1.00
Diminished:BlockOrder	0.45	0.29	-0.13	1.02	7464	1.00
Augmented:BlockOrder	0.60	0.33	-0.04	1.27	9491	1.00

Significant population-level effects for Model 3.2 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ ). The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

Significant in this model, in the positive direction, are MusSoph, Previous and SPCS. Minor, diminished and augmented triads all received significantly lower ratings than major triads, the degree to which increases significantly with musical sophistication. Minor triads also received significantly higher ratings than diminished and augmented triads. Plots of the conditional effects of Triad and MusSoph:Triad are included below, in Figures 3.11 and 3.12. The positive interaction between MusSoph and SPCS is also significant. Significant interactions with MusSoph confirm the third hypothesis. Finally BlockOrder:Minor comes in (only just) as significant. A plot of the conditional effect of BlockOrder:Triad is shown in Figure ?? in Appendix C. Though the differences between major and diminished or augmented triads are not significant, the plot suggests that when the probe triad block was heard

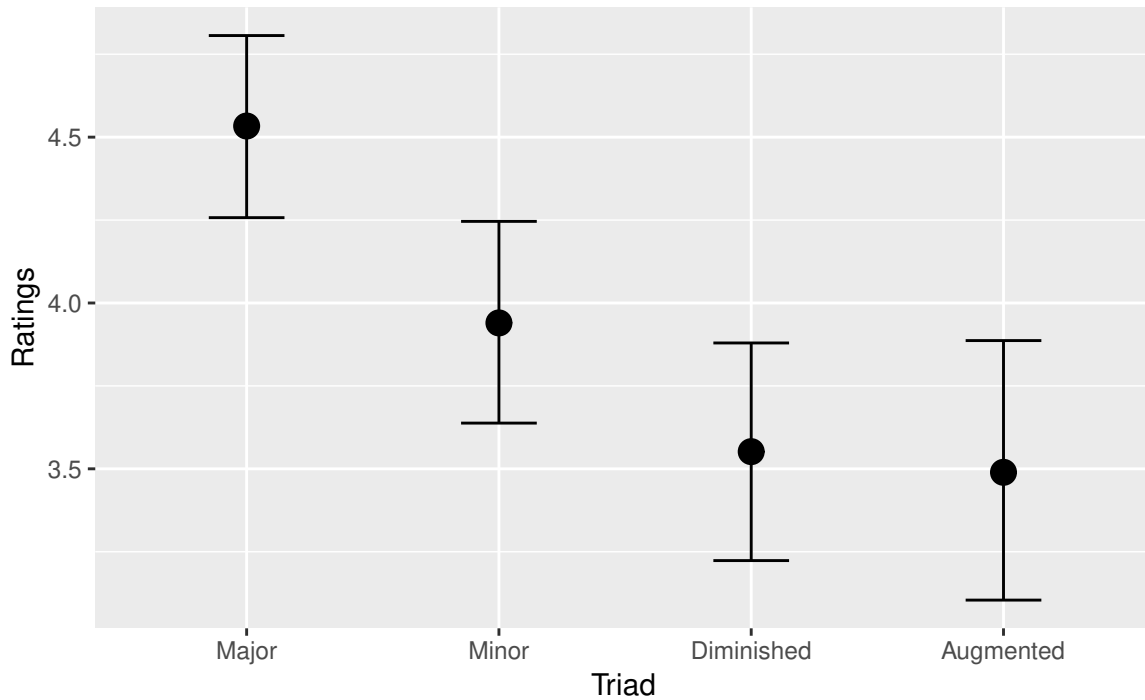


FIGURE 3.11: Conditional effect of Triad for Model 3.2

before the probe tone block, major triads were given a clearly higher rating, but all ratings of all other triad types cross error bars. Presently we have no explanation for this.

Table 3.7 shows the results of hypothesis tests on all significant effects in the model. All hypotheses for the significant effects have very strong evidence to support them.

TABLE 3.7: Evidence ratios for all significant effects of Model 3.2

Hypothesis	evidence ratio
SPCS > 0	> 11999
MusSoph > 0	163.38
Previous > 0	> 11999
Maj > Min	> 11999
Maj > Dim	> 11999
Maj > Aug	> 11999
Min > Dim	314.79
Min > Aug	199
MusSoph:SPCS > 0	704.88
MusSoph:Maj > MusSoph:Min	665.67
MusSoph:Maj > MusSoph:Dim	64.93
MusSoph:Maj > MusSoph:Aug	922.08
BlockOrder:Min > BlockOrder:Maj	51.86



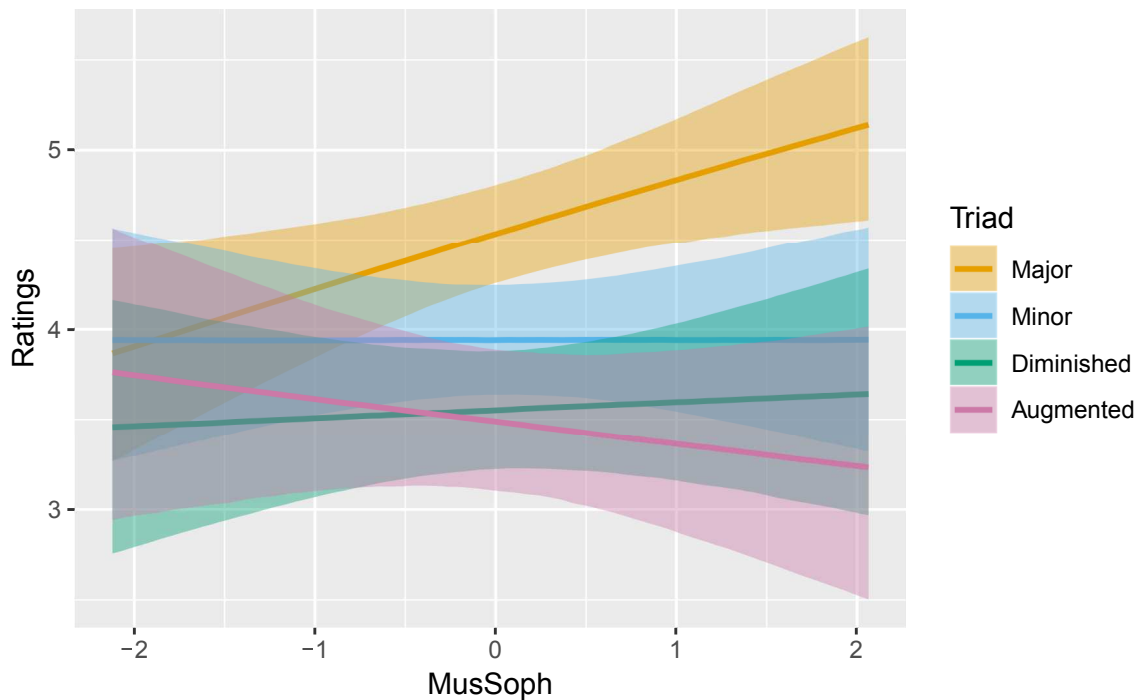


FIGURE 3.12: Conditional effect of MusSoph:Triad for Model 3.2

Whereas in Model 3.1 MusSoph had a significantly more strongly positive effect for minor and for diminished triads than for major triads, for these scales the effect is significant in the opposite direction (the effect of MusSoph was significantly larger for diminished triads than for augmented or minor triads for Model 3.1, both differences insignificant in this model).

A posterior predictive check for this model shows it to be a valid model for the data. The pseudo- $R^2$  value for this model, at .51, is higher than that of Model 3.1, and very similar to the pseudo- $R^2$  value for the probe tone model for this experiment 2, though it included four additional scales.

Model 3.2 is compared via cross-validation to four different models with the results shown in Table 3.8. First, confirming H2, it is compared to a model differing by the absence of both SPCS and Triad as predictors and found to perform significantly better. It is also found to perform significantly better than models that differ only by the absence of either SPCS or Triad. Finally, Model 3.2 is compared to a model that differs only by the addition of Inversion, wherein no significant difference is found.

TABLE 3.8: LOOIC comparisons for Experiment 2 Bayesian ordinal mixed effects models

Model compared to Model 3.2	Model – Model 3.2	LOOIC	SE	Signif
– Triad – SPCS		610.8	63.0	yes
– Triad		479.2	57.0	yes
– SPCS		119.0	26.2	yes
– SPCS + ScaleTone		4.2	18.2	no
+ Inversion		31.0	10.0	no
Model compared to Model 3.3	Model – Model 3.3	LOOIC	SE	Signif
– SPCS + ScaleTone		93.0	24.4	yes

### Combined analysis: Model 3.3

Given that the method differs only by which probes and scales are used in stimulus, and the models differ only by the inclusion of Prevalence only in the confirmatory model for Experiment 1 (Model 3.1), the probe triad data from Experiments 1 and 2 can be combined and a model equivalent to Model 3.2 run for this combined data set. This model – Model 3.3, whose results are summarized in Table 3.9 below – is more informative than those above as it includes the most wide-ranging data and number of observations. Table 3.10 shows the evidence ratios for all significant effects in Model 3.3 and Table C.8, in Appendix C, shows all population-level effects.

In Model 3.3 we see more significant effects than in Models 1 and 2 due to the fact that this model includes more data. Excluding the interactions between MusSoph and Triad, all effects significant in these models are significant in Model 3.3, apart from Recency, (evid. ratio 13.9), and any interactions of BlockOrder with Triad (as well as Prevalence and MusSoph:Prevalence, as these is no Prevalence in this model).

Considering interactions between MusSoph and Triad, which differed for Models 3.1 and 3.2, we see a mixture of both: As in Model 3.1, the effect of MusSoph is more strongly positive for diminished than for augmented triad. Unlike in Model 3.1, and like in Model 3.2, the effect of MusSoph is more strongly positive for major than minor triads, and like in Model 3.2, more strongly positive also than for augmented triads. Additionally, it is more strongly positive for minor triads than for augmented triads in Model 3.3. Figure 3.13 – the conditional effect of MusSoph:TriadNo shows that, closer to what we might expect than in Model 3.1, as MusSoph increases, major triads get increasingly higher ratings, and augmented triads get increasingly lower

ratings, with minor and diminished triads getting slightly higher ratings (MusSoph has a significant positive effect on ratings).

TABLE 3.9: Model 3.3 significant population-level effects

Effect	Estimate	Est. Error	l-95% CI	u-95% CI	Eff. Sample	Rhat
Intercept[1]	-3.40	0.11	-3.61	-3.19	4417	1.00
Intercept[2]	-2.07	0.10	-2.27	-1.86	4281	1.00
Intercept[3]	-0.89	0.10	-1.10	-0.69	4284	1.00
Intercept[4]	-0.18	0.10	-0.39	0.02	4219	1.00
Intercept[5]	1.00	0.10	0.80	1.20	4234	1.00
Intercept[6]	2.48	0.10	2.28	2.69	4443	1.00
MusSoph	0.19	0.09	0.02	0.37	3584	1.00
Height	-0.08	0.18	-0.42	0.27	6805	1.00
RelHeight	-0.00	0.18	-0.34	0.34	7108	1.00
Previous	0.29	0.04	0.21	0.37	7266	1.00
Recency	0.09	0.06	-0.02	0.20	10517	1.00
SPCS	0.37	0.04	0.28	0.45	7189	1.00
Count	-0.01	0.05	-0.11	0.10	10996	1.00
Minor	-0.70	0.09	-0.87	-0.52	6171	1.00
Diminished	-1.10	0.12	-1.32	-0.87	5937	1.00
Augmented	-1.43	0.14	-1.72	-1.15	7103	1.00
TrialNo	0.08	0.05	-0.03	0.18	8636	1.00
InContTrialNo	-0.09	0.04	-0.17	-0.01	10495	1.00
BlockOrder	-0.26	0.19	-0.64	0.11	3616	1.00
MusSoph:SPCS	0.22	0.04	0.15	0.31	6923	1.00
MusSoph:Count	0.08	0.03	0.01	0.15	11385	1.00
MusSoph:Minor	-0.13	0.08	-0.30	0.03	5485	1.00
MusSoph:Diminished	-0.08	0.11	-0.30	0.13	5120	1.00
MusSoph:Augmented	-0.40	0.12	-0.64	-0.16	6293	1.00
Height:TrialNo	0.35	0.17	0.01	0.69	6916	1.00
RelHeight:TrialNo	-0.37	0.17	-0.70	-0.03	6911	1.00
Minor:TrialNo	-0.10	0.05	-0.19	0.00	10542	1.00
Diminished:TrialNo	-0.12	0.05	-0.23	-0.02	11119	1.00
Augmented:TrialNo	-0.03	0.08	-0.19	0.13	12966	1.00
Count:InContTrialNo	-0.05	0.02	-0.09	-0.01	14806	1.00
Minor:InContTrialNo	0.09	0.04	0.01	0.17	12259	1.00
Diminished:InContTrialNo	0.12	0.05	0.03	0.21	13281	1.00
Augmented:InContTrialNo	0.09	0.07	-0.04	0.22	15210	1.00
Count:BlockOrder	-0.27	0.09	-0.44	-0.10	10196	1.00

Significant population-level effects for Model 3.3 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

TABLE 3.10: Evidence ratios for all significant effects of Model 3.1

Hypothesis		evidence ratio	strength
SPCS	>0	> 11999	very strong
MusSoph	>0	81.19	very strong
Previous	>0	> 11999	very strong
Maj	>Min	> 11999	very strong
Maj	>Dim	> 11999	very strong
Maj	>Aug	> 11999	very strong
Min	>Dim	> 11999	very strong
Min	>Aug	> 11999	very strong
Dim	>Aug	138.53	very strong
MusSoph:Maj	> MusSoph:Min	19.3	strong
MusSoph:Maj	> MusSoph:Aug	1713.29	very strong
MusSoph:Min	> MusSoph:Aug	92.75	very strong
MusSoph:Dim	> MusSoph:Aug	271.73	very strong
TrialNo:Maj	> TrialNo:Min	33.78	very strong
TrialNo:Maj	> TrialNo:Dim	85.33	very strong
InContTrialNo:Min	> InContTrialNo:Maj	66.04	very strong
InContTrialNo:Dim	> InContTrialNo:Maj	362.64	very strong
InContTrialNo:Aug	> InContTrialNo:Maj	10.17	strong
MusSoph:SPCS	>0	> 11999	very strong
MusSoph:Count	>0	110.11	very strong
Height: TrialNo	>0	45.51	very strong
RelHeight: TrialNo	<0	58.41	very strong
Count: InContTrialNo	<0	180.82	very strong
Count: BlockOrder	<0	856.14	very strong

We also see effects significant in Model 3.3 that though present in the same direction, were not significant in either Model 3.1 or Model 3.2: Diminished triads are rated significantly higher than augmented triads. The effect of Trial Number is more strongly positive for major than for minor and diminished triads. Within a context scale, however, the positive effect is weaker for major triads than for minor, diminished or augmented triads. Figures 3.14, 3.15 and 3.16, show the conditional effects of Triad, TrialNo:Triad and InContTrialNo:Triad.

Finally, the interaction between Height and Trial number is also significant in this model.

### Exploratory analysis

Models 3.2 and 3.3 are compared to models equivalent, but with ScaleTone replacing SPCS. Regarding Model 3.2, no significant difference was found. In contrast, Model 3.3 significantly outperformed its related model. Both comparisons are shown in Table 3.8. ScaleTone and MusSoph:ScaleTone are significant positive effects in

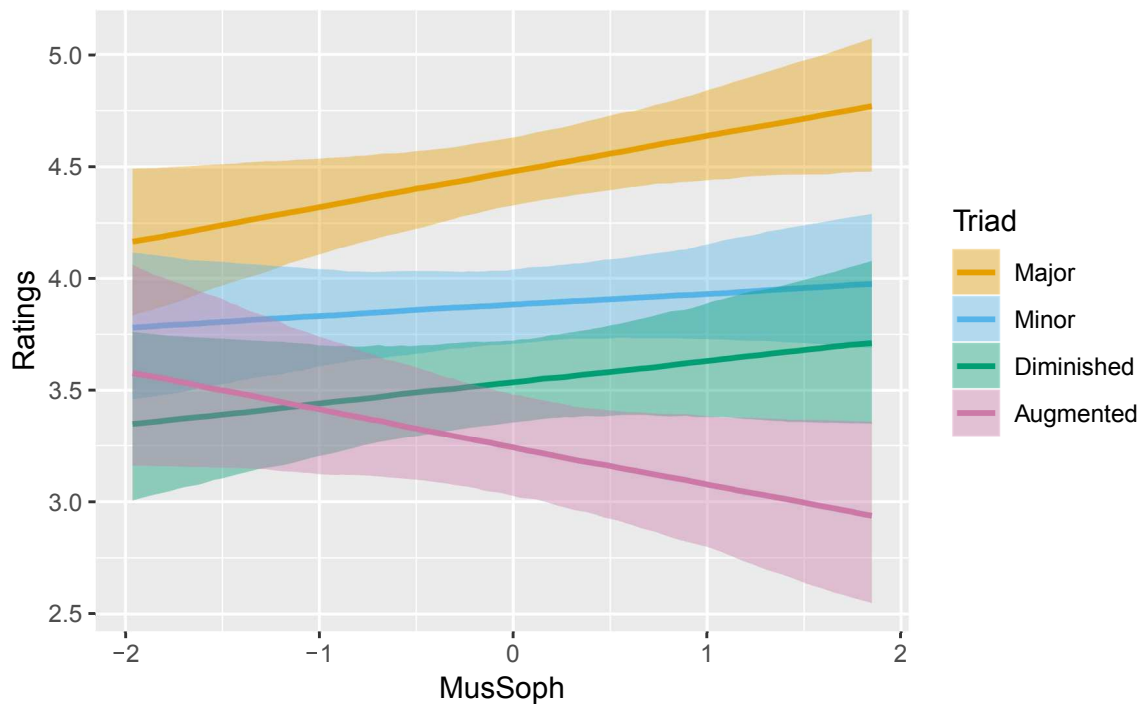


FIGURE 3.13: Conditional effect of MusSoph:Triad for Model 3.3

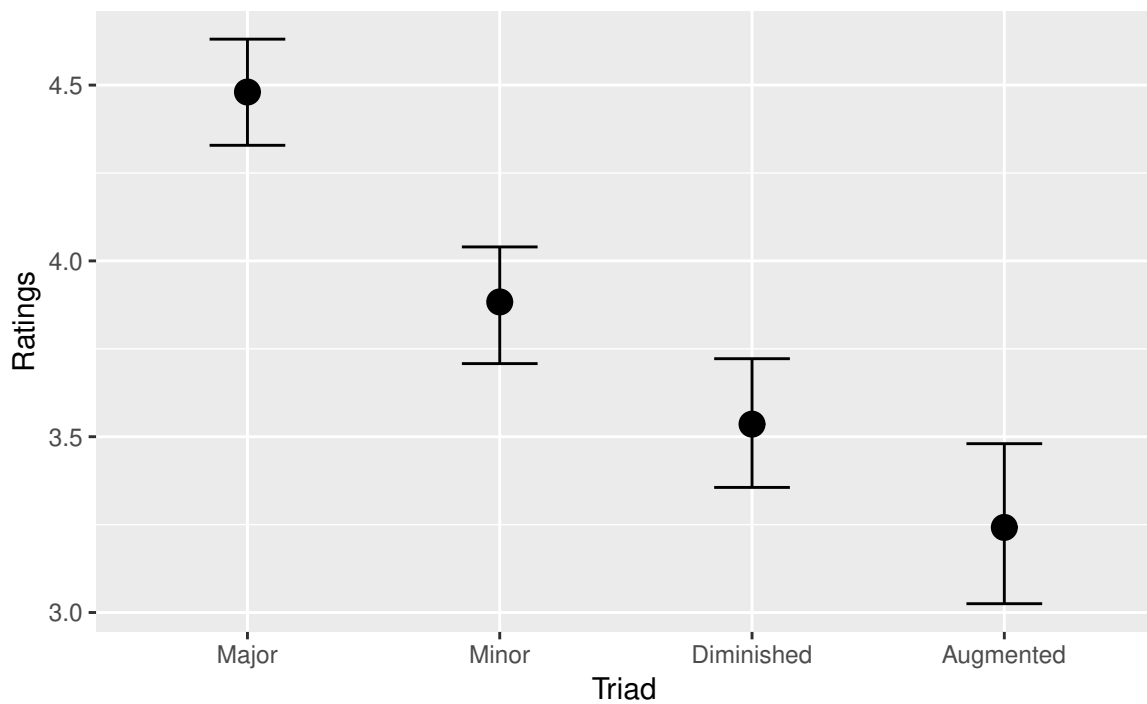


FIGURE 3.14: Conditional effect of Triad for Model 3.3

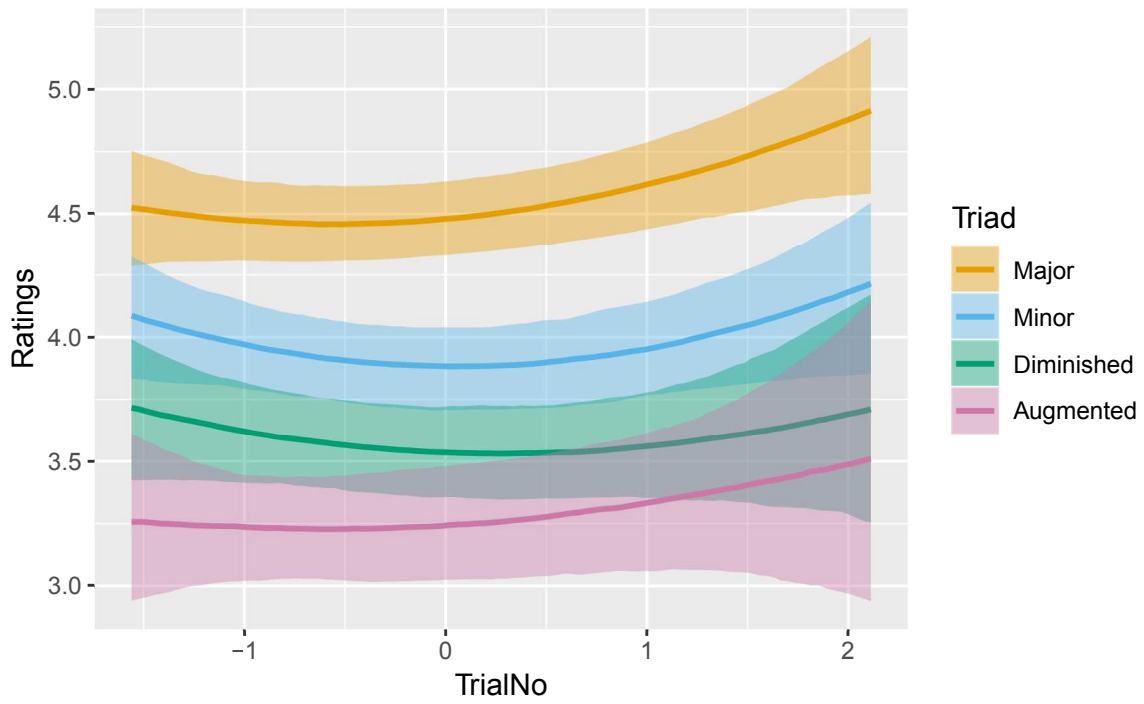


FIGURE 3.15: Conditional effect of TrialNumber:Triad for Model 3.3

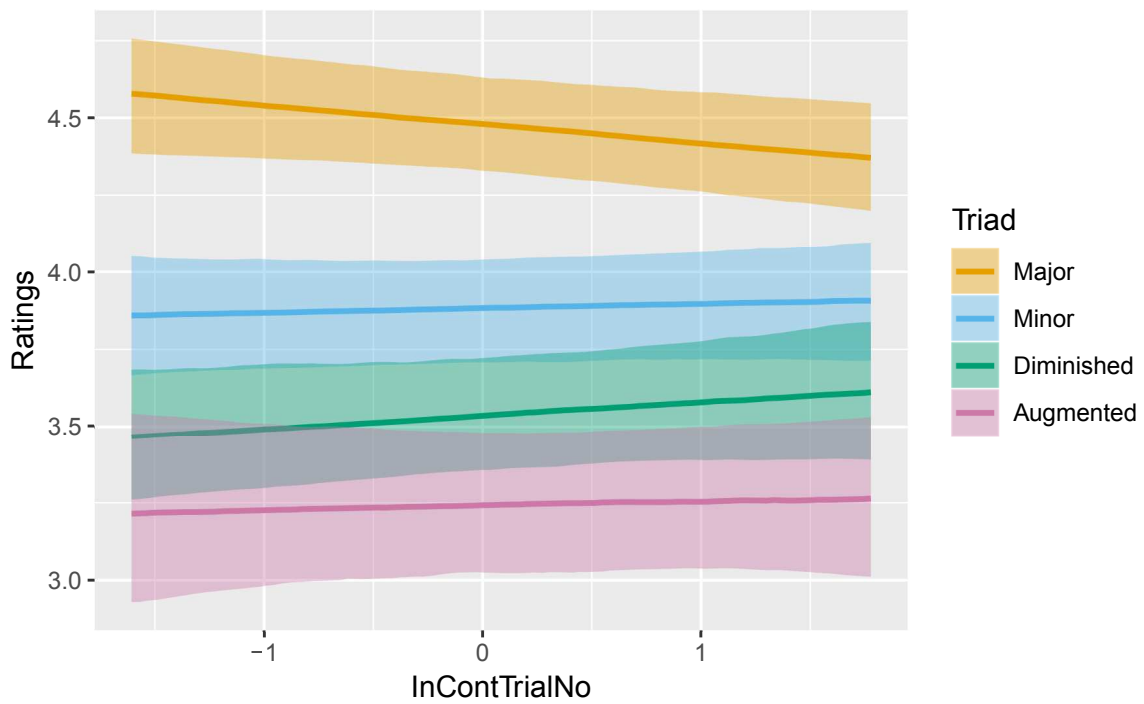


FIGURE 3.16: Conditional effect of InContTrialNo:Triad for Model 3.3

both these models (see Tables C.7 and C.9 in Appendix C for significant population-level effects of the models, along with intercepts and associated conditional effects). The results of the comparison for Model 3.3 reflects that of the exploratory analyses of Experiment 1 when Prevalence is not included, however for the scales of Experiment 2 it seems SPCS is no better than ScaleTone.

### 3.4.3 Discussion

The descriptive model hints at the tonic function belonging perceptually to a single chord in both scales – the chord that serves as a theoretical tonic – as in the scales of Experiment 1. This was not the case for the probe tones (Chapter 2), where – although single perceptual tonics were observed for the three scales of Experiment 1 – such tonics were not observed or hinted at for the harmonic major and double harmonic scales. This suggested that the use of triads may aid the possibility for a perceived tonic in a scale, i.e., that harmonic tonality may be a stronger form of tonality.

Tertian triads in these unfamiliar scales include triads that are not Major, Minor, Diminished and Augmented. These triads are not used as probes as they are less commonly used in Western music, and only exist in a small number in the full set of scales tested here. For this data set the fact that SPCS is still a significant positive effect, as a conditional main effect and as an interaction with MusSoph, suggests SPCS is a generalizable measure of tonal fit; however, a simple short term memory model – ScaleTone, was found in some exploratory analyses to be equally effective a predictor.

We may interpret from Figure 3.13 that musical sophistication increases the ability to tell apart different triads, or perhaps just that dissonant triads are given increasingly lower ratings, and consonant triads increasingly higher. The effect of consonance/dissonance increases with music training, either as musicians develop their audition skills, or learn to experience or identify consonance through explicit or implicit learning. Superimposing different aspects of statistical learning might provide a testable explanation for the fact that major triads are given higher ratings across trial number, but lower across trial number within a context. They are, however, both very small effects.

As in Experiment 1, the performance of Model 3.2 as significantly above its associated null confirms H2. H3 is confirmed by the significance in Model 3.2 of interactions with MusSoph. Models 3.2 and 3.3 are significantly weakened by the removal of effects of SPCS. We can be fairly certain that SPCS measures an effect important in the cognition of harmonic tonality and that this effect is enhanced by musical sophistication, possibly due to training of audition skills.

### 3.5 Conclusion

In Experiment 1 participants were played randomly generated, isochronous melodies using notes from the diatonic, harmonic minor and jazz minor scales. Half the participants rated the goodness-of-fit of probe tones played after these scales, and the other half rated their stability. Though differing significantly for the tonic and leading tone triads of the jazz minor scale, supporting our first hypothesis, we found that, overall, perceived goodness-of-fit can be considered equivalent to perceived stability. Experiment 2 differed from Experiment 1 in the scales used for context melodies – the harmonic major and double harmonic scales. For this experiment we hypothesised (H1) that given the lower level of familiarity with the scales used for the context stimulus, we would not see significant difference between fit and stability ratings for any pitch-classes. Our results supported this hypothesis.

To remind the reader of our second and third hypotheses, they were:

H2: Perceived goodness-of-fit and perceived stability of probe triads may be modelled by the SPCS between the pitches of the context and the probe, and the consonance/dissonance of the probe and, for Experiment 1, the statistical prevalence in Western music of the probe within the context scale.

H3: Significant interaction effects exist between musical sophistication and other predictors in such a model.

Considering these hypotheses, for our first data set, of familiar scales (Experiment 1), our second data set of less familiar scales (Experiment 2), and for our combined data set we have shown that SPCS is a significant predictor of fit and stability ratings. Removing SPCS, Prevalence and Triad from Model 3.1 and SPCS and Triad from Model 3.2 and Model 3 significantly weakens the models. Thus, our second and third hypotheses are confirmed for Experiments 1 and 2. Also significant in all models is Triad, from which we recover a reinforcement of the observed



relative consonance of triads, namely, augmented < diminished < minor < major. Though they differ between models, in Model 3.3, which models all of our data, we find that overall more musically sophisticated participants rate more consonant triads higher, and more dissonant triads lower, whereas for participants with lower musical sophistication there is little or no difference.

Though the scales used in Experiment 2 were unfamiliar, they nonetheless do occur in Western common-practice music, and accordingly we cannot rule out effects of familiarity in our ratings, even if we cannot test for them due to the lack of a representative corpus. Considering, however, the existence of cases where models of prevalence do not easily apply (for example, using scales for which there is no associated corpus data) we can already be confident that SPCS is a more widely applicable model than Prevalence, where, unlike Prevalence and top down models, SPCS can be immediately generalized to novel stimuli. We note that it may be possible to build a model of statistical learning using IDyOM (Pearce, 2005), however since such a model is still no more widely applicable we have not pursued this option. Dean et al. (2019) also demonstrated that SPCS can predict perceived change in sound-based as well as note-based music, strengthening further the argument for its usefulness as a psychoacoustic measure. It would seem wise still to test both for this capacity for generalisation in a context in which effects of familiarity can be further diminished. Accordingly, in Chapter 6 we use truly novel (microtonal) scales for our context stimulus. Before we can test for the perceived stability of RPCs and triads in such scales, we first test for the intrinsic stability of all triads – stability due to the triads isolated from any context – of 22-TET, the tuning system from which we obtain our scales, detailed in Chapter 4. In order to arrive at our selection of scales to be used as context stimulus we first complete a distributional analysis of all possible seven-tone scales of 22-TET in terms of their values of scale features, detailed in Chapter 5.

## Chapter 4

# Experiment 3: Intrinsic stability of the triads of 22-TET

### 4.1 Introduction

The remainder of the thesis concerns triads and scales in 22-TET. This chapter tests for the stabilities of all triads of 22-TET; Chapter 5 looks into features of all seven-note scale of 22-TET, and Chapter 6 tests for the perceived stability of triads and RPCs after the context of a selection of 22-TET scales.

In Experiment 3, detailed in this chapter, we test the perceived stability of all 210 possible 22-TET triads (assuming transpositional equivalence) within an octave after a context of a randomly ordered, isochronous sounding of each note of 22-TET. Unlike all the previous experiments the context is not a scale, but a complete set of the available notes of 22-TET. Consequently, the relative perceived stability of the 22-TET triads provides a measure of stability intrinsic to the triads themselves. After a simple linear model using the triad type to model perceived stability, a Bayesian mixed effect model is run, controlling for effects of the context. A *leave-one-out cross validation information criterion*, or *LOOIC* comparison of Bayesian ordinal mixed effect models this model against one that differs only by the absence of an effect of triad type confirms the existence of such intrinsic stability.

This provides a background from which we test the effect of different 22-TET scales as context on the perceived stability of triads in Experiment 5, as detailed in Chapter 6. From the use of different 22-TET scales we can test for the effect of SPCS on the perceived stability of triads. Chapter 6 also outlines the results of Experiment 4, which considers the perceived stability of tones after the context of the same microtonal scales. Chapter 5 details a distributional analysis used to choose

the set of scales tested in Experiments 4 and 5. This involves a cluster analysis of the scales in terms of the values of many features. Most relate to their structure, but three of these features relate to the perceived stability of the tertian triads available in the scale, from the results of this experiment. This experiment – Experiment 3 – concerning the perceived stability of all triads of 22-TET was pre-registered at Open Science as part of the pre-registration challenge, available at <https://osf.io/t9xz6>.

## 4.2 22-TET

The choice of 22-TET as a tuning system for our study is not arbitrary. Though the study of 19-TET is more thorough and dates further back, 22-TET garners much modern interest from theorists and musicians. The first theoretical discussion of 22-TET within the context of Western music was given by Bosanquet (1878), though earlier work erroneously describes the Indian sruti system as 22-TET (the 22 sruti of Indian music are known today to be spaced un-equally within an octave) (Monzo, n.d.). Würschmidt (1921) advocated for the use 22-TET for the future, to follow the extensive of 19-TET. 22-TET is discussed along with other ETs in later explorations of tuning systems such as those by Barbour in 1951 (Barbour, 2004), and Blackwood (1985). The first musical use of 22-TET was provided by prolific microtonal musician Iver Darreg in the 1960s, and in 1980, pianist and composer Moshe Cotel briefly explored the tuning (Monzo, n.d.). In the same year pianist, composer and professor of music Easley Blackwood released his seminal *Twelve Microtonal Etudes for Electronic Music Media*, comprising an etude in each ET from 13-24 (Blackwood, 1980). Paul Erlich's exploration of 22-TET began in 1993 (Monzo, n.d.), and lead to the publication of his 1998 paper *Tuning, Tonality, and Twenty-Two Tone Equal Temperament* (Erlich, 1998), in which he advocated for the use of a 10-note scale in 22-TET as a way to continue the evolution of tonal-harmonic systems in Western music. This paper, along with his composition *TIBIA* (<http://www.tallkite.com/words/Tibia.mp3>) was widely celebrated in online microtonal communities and inspired many other composers to explore the tuning. Today, a number of contemporary musicians and bands embrace 22-TET in their music, including ILEVENS (<https://www.youtube.com/watch?v=5KsrnvZjvWo>), Redrick Sultan (<https://redricksultanband.bandcamp.com/track/recurring-mimosa-2>), Sevish (<https://www.youtube.com/watch?v=oNPCiBY5IZ8>),

Brendyn Byrnes (<https://brendanbyrnes.bandcamp.com/track/22-edo-guitar-etude>) and Jacob Barton (<https://soundcloud.com/metaclown/couples-therapy>).

The ability of 22-TET to approximate frequency ratios with odd numbers up to 11 (11-odd-limit frequency ratios) remarkably well for its size is celebrated (“22edo”, 2020). Partch described such a collection of ratios as the ‘11-limit tonality diamond’, and considers them to be new possibilities for consonances in Western Music (Partch, 1949). Though 12-TET very accurately approximates frequency ratios with odd numbers up to 5, its approximations of 11-odd-limit intervals comprising numerators or denominators of 7 and 11 are relatively poor (“12edo”, 2021). Though 19-TET and 31-TET, for example, also provide closer approximation to 11-odd-limit frequency ratios, 22-TET is the smallest ET to approximate them consistently (“22edo”, 2020). This means that its best approximation of any 11-odd-limit ratio resulting from the multiplication or division of two 11-odd-limit ratios A and B is equivalent to the interval of the ET resulting from the addition or subtraction of its closest approximations of the 11-odd-limit intervals A and B (“Consistent”, 2019). Given that 12-TET’s best approximation of the interval  $11/9$  ( $\sim 347c$ ) is 3 degrees, and the best approximation of the interval  $9/8$  ( $\sim 204c$ ) is 2 degrees, whereas the best approximation of the interval  $11/9 \times 9/8 = 11/8$  ( $\sim 551c$ ) is 6 degrees, which does not equal  $2 + 3$  degrees, 12-TET is not consistent in the 11-odd-limit. Figures 4.1a and 4.1b and provide a visualization of 12-TET’s and 22-TET’s approximation (respectively) of 11-integer-limit frequency ratios. Only ratios that are approximated to within one third of a degree of the ET are displayed in the figure.

Criticism of 22-TET often highlights its inability to support meantone temperament (defined in 11), the tonal-harmonic system underpinning a large proportion of Western “common practice” music that is supported by 12-TET as well as 19-TET and 31-TET (Blackwood, 1985). Given that we do not require the ability to support familiar tonal systems in our tuning, however, this criticism is irrelevant in this context. Further, the inability for 22-TET to support meantone temperament also aids in the disentanglement of tonal-harmonic features of the diatonic scale explored in Chapter 5.

Finally, the larger size of ETs which more accurately approximate 11-limit consonances, such as 31-TET, 41-TET, 46-TET and 72-TET makes it harder to completely examine their resources, as we have been able to do with 22-TET in this chapter and in chapter 5. Compared to the 55 possible triads and 66 possible 7-note scale of 12-TET and the 210 possible triads and 7752 possible 7-note scales of 22-TET, there

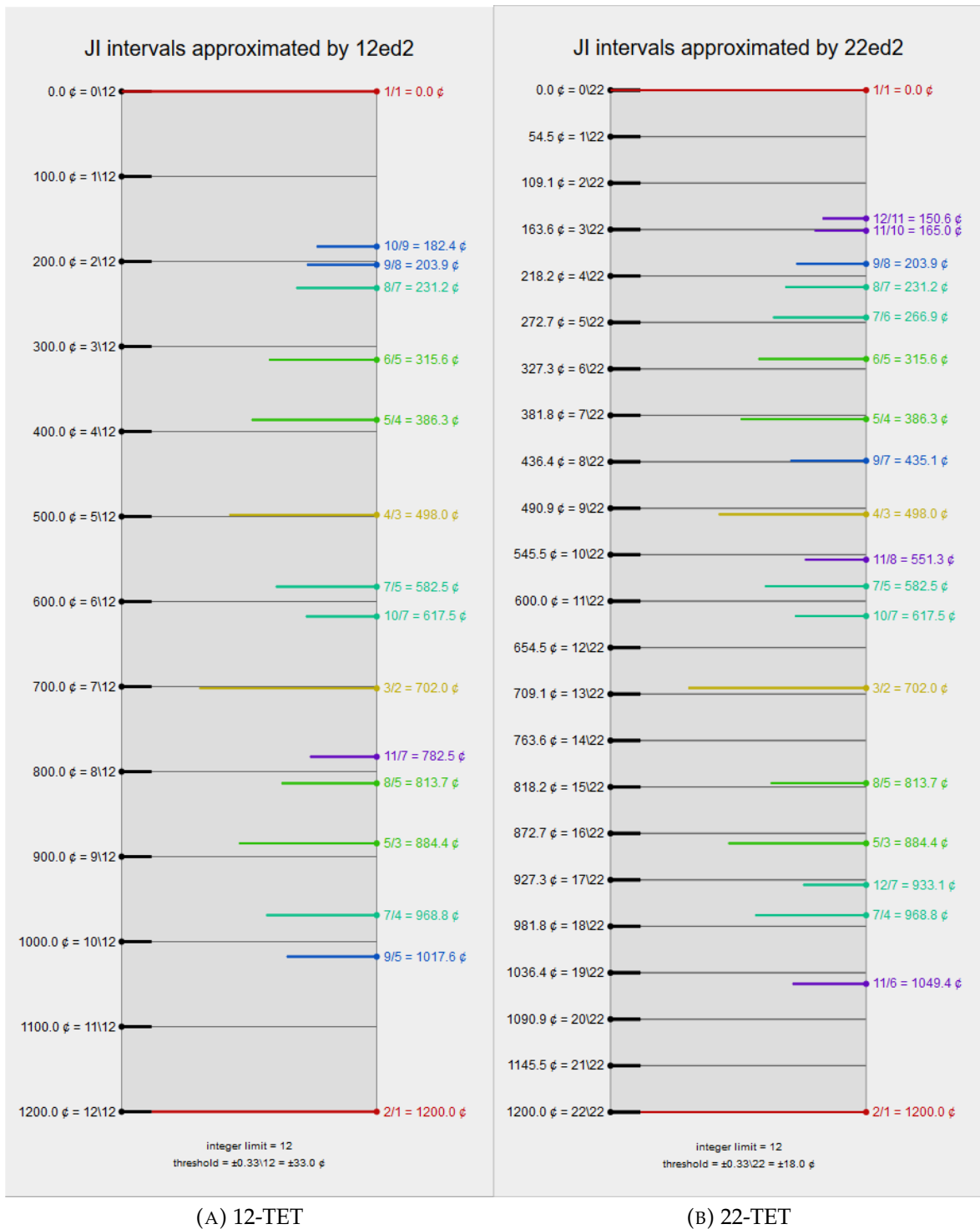


FIGURE 4.1: A visualisation of 12-TETs and 22-TETs approximation of 11-integer-limit frequency ratios

are 435 possible triads and 84825 possible 7-note scales in 31-TET. 22-TET provides the possibility for novel tonal-harmonic systems using largely unexplored 11-odd-limit consonances, whilst still providing the familiar consonances, with a minimal increase of cardinality.

### 4.3 Hypothesis

The hypotheses as preregistered, to be tested in 22-TET, read:

H1: Triad types possess intrinsic stability.

H2: The intrinsic stability of triads may be modelled by an additive combination of models of roughness, harmonicity, spectral entropy, harmonic entropy, and familiarity.

H2 is not tested in this dissertation; it is left for future work.

### 4.4 Method

#### 4.4.1 Participants

Twenty-four musicians (participants who reported having received 5 or more years of music experience) and 36 non-musicians were recruited for the experiment. Non-musician participants were university students (mostly first-year) recruited through Western Sydney University School of Psychology and Social Science's SONA system and received credit points towards their degrees for their participation. Musician participants were recruited via personal connection and received a \$30 reimbursement for their time and travel to the university campus. All participants demonstrated normal hearing capabilities. Fourteen (musician) participants reported having received 10 or more years of musical training and two reported having absolute pitch. Participants had a mean age of 26 years, with a SD of 10.9 years. Of the 24 musicians, 16 were female, and of the 36 non-musicians, 26 were female. This research was approved by the Western Sydney University Human Research Ethics Committee under the number H11908.

### 4.4.2 Stimuli and Procedure

The stimuli differ from those of Experiments 1 and 2 only by the context scales and the probes.

Rather than context stimuli consisting of soundings of each note of a number of scales, in this experiment each pitch of one octave of 22-TET is sounded a single time.

For each trial, participants were asked to rate the ‘stability of the final musical sound given the sounding of the context melody’, entering their ratings on a 7-point Likert scale presented on-screen horizontally, as in Experiments 1 and 2, detailed in Section 2.3.1. Participants were also informed that ‘a musical sound is considered to be stable if it does not need to move (resolve) to another music sound’.

The possible number of triads of 22-TET whose range does not exceed an octave is  $21 \times 20/2 = 210$ . Should each triad be heard twice by each participant that would mean 420 experimental trials, which would be much too long for a single sitting. Accordingly, the triads are split across 30 pairs of participants. The first in each pair hears a random selection of triads twice each, and the second hears the other half. The experimental trials are preceded by 6 practice trials. Half way through the experiment participants completed a survey including the Goldsmith MSI Questionnaire, in order to obtain an index for musical sophistication to be used as a variable in analysis. Additional demographic questions followed the Goldsmith MSI Questionnaire to facilitate future analysis of possible effects of enculturation, though these are not analysed here.

## 4.5 Analysis

A Bayesian ordinal (cumulative logit) mixed effects model is run, similar to that of the triad models of Experiments 1 and 2, but for the absence of BlockOrder and Task, since this experiment is not blocked only stability is rated by participants. The model’s effects are as follows:

- *SPCS*: The spectral pitch class similarity of the aggregated pitches in the context and the probe.
- *Triad*: The triad type of the probe. Discussed in more detail below.
- *Recency*: Coded as 1 when one of the pitches of the probe matches the final pitch of the context, and 0 otherwise.

- *Primacy*: Coded as 1 when one of the pitches of the probe matches the initial pitch of the context, and 0 otherwise.
- *MelCont*: The melodic continuity of the probe from the context – the maximum number of consecutive intervals in a single direction that can be traced back from any of the three pitches of the probe (may take integer values from 1 to 7, given the one octave range).
- *Previous*: The rating given to the previous trial.
- *Count*: The sum of the number of occurrences in the experimental stimulus of the three pitches of the probe at the time of the trial.
- *RelHeight*: The sum of the pitch heights of the three pitches of the probe relative to its context (measured in semitones above the lowest pitch of the context. May take integer values from 3 to 60).
- *RelHeight<sup>2</sup>*: The relative pitch height of the probe squared.
- *Height*: The sum of the pitch heights of the pitches of the probe.
- *Height<sup>2</sup>*: The square of the sum of the pitch heights of the pitches of the probe.
- *TrialNo*: Trial number.
- *TrialNo<sup>2</sup>*: Trial number squared.
- *MusSoph*: The musical sophistication of the participant, as measured by the Goldsmith Musical Sophistication Index.

As in Experiments 1 and 2, we intended to split the variables into two groups – those which represent a feature of the stimulus, and those which may affect the relative influence of such a feature on the participant’s rating. In such a division each variable of the second group would interact with each variable of the first group. The second group consists of *MusSoph*, *TrialNo*, *TrialNo<sup>2</sup>*; the first group comprises the remaining effects. However, *Triad*, as in Experiments 1 and 2, is categorical, and there are 210 possible triads within an octave of 22-TET that were tested as probes. The model became unfeasibly large, and so interactions with *Triad* were not included.

As stated in the preregistration, following the recommendation of Barr et al. (2013) we intended to run the model with the maximal random effects structure driven by the design of the experiment, including random effects on participants with respect to the first group of variables, namely: *Triad*, *SPCS*, *Primacy*, *Recency*, *Melodic Continuity*, *Count*, *Height*, *Relative Height*, *Height<sup>2</sup>* and *Relative Height<sup>2</sup>*; and their correlations. However, again, considering the number of categories of



triad the model became infeasible large, and so random effects were initially not included for Triad. Each of the 210 triads was represented by a separate category, dummy coded against the most closely voiced triad possible in 22-TET – a stack of 3 single-degree intervals, represented as [0,1,2], meaning the triad with the second note 1 degree above the first, and the third note 2 degrees above the first. Though in Experiments 1 and 2 our models performed better with inversional equivalence assumed, we cannot assume the same to be true for microtonal scales. Indeed, Mathews et al. (1988) found that for triads in Bohlen-Pierce tuning, inversional equivalence was not observed, however, unlike 12-TET and 22-TET, the Bohlen-Pierce scale does not observe octave equivalence, repeating instead at the interval corresponding to a frequency ratio of 3/1 (rather than 2/1).

To test whether the assumption of inversional equivalence improves our model of triads of 22-TET, we run a model in which inversional equivalence is assumed, where there are therefore 70, rather than 210 triad categories (this model also did not include random effects on participants with respect to triad). A LOOIC comparison shows the model assuming inversional equivalence to be insignificantly stronger, with a difference in LOOIC of 28.30, within the SE of 33.38. As it is much simpler and not significantly worse, we assume inversional equivalence (as well as transpositional) for the categories of triad in our model. Given the assumption of inversional equivalence, our model is now much smaller, and random effects on participants with respect to Triad is included as intended in the preregistration, along with SPCS, Primacy, Recency, Melodic Continuity, Count, Height, Relative Height, Height<sup>2</sup> and Relative Height<sup>2</sup>; and their correlations.

The context stimulus for each trial includes all notes of 22-TET within an octave, which are, by definition, evenly spaced. The context stimulus for each trial therefore contains the same (equally spaced) pitch class content. Accordingly, once all other effects are controlled for, the stability of the triad given the context of the scale must be due to the triad itself, and not the context, and therefore represents the intrinsic stability of the triad. The model was too large to be used in a LOOIC comparison, so H1 is tested instead with a model identical to Model 4, but where no interactions with Triad are included, and random effects on participants with respect to Triad are not included. This model was compared via LOOIC comparison of the above model with an associated null – a model identical but for the absence of an effect of Triad. A comparison is said to be significant, as in Experiments 1 and 2, when the difference in LOOIC is greater than twice the associated SE.

H2 is not tested here, as it was not necessary for the primary investigations considered in this thesis, and will be tested in the future.

In this model, as the number of triads is much larger than in the models of Experiments 1 and 2 – 70 compared to 4 – each triad is modelled as a value of one for its category, and values of 0 for all other categories. Binary effects, then, greatly outnumber continuous effects in the model. Accordingly, we scale the continuous variables such that the standard deviation is  $1/2$ , rather than 1, to line up with the triad category effects, which take the values 0 or 1. This is in line with the recommendations of Gelman and Hill (2006). A

$\text{student\_t}(3, 0, 2.5)$

(a  $t$ -distribution with 3 degrees of freedom, with mean of 0, scaled by 2.5) is used again as a weakly informative prior.

## 4.6 Results

### 4.6.1 Descriptive analysis

A linear model is first run, using Triad to predict perceived stability. Triads in 22-TET do not have standard names at this stage, as exploration into the tuning system is still in its early stages. In this chapter triads are labelled in their *normal form* – expressed in the inversion such that the outer interval is first minimized, and then, in the case that two inversions share the same outer interval, the lower interval is minimized – with the bottom note on pitch class 0. Normal form is typically applied only to 12-TET, in which the pitch classes of each note are expressed as degrees of 12-TET from the tonic. In this paper the notes of the triad are expressed as degrees of 22-TET above the bottom note. For example, 22-TET's best approximation of the *classic major triad* represented by the frequency ratios 4:5:6 (or the fourth, fifth, and sixths overtones of the harmonic series) includes a major third interval (approximating  $5/4$ ) 7 degrees ( $\sim 382c$  cents – hundredths of a 12-TET semitone – above the root)<sup>1</sup> and a minor third interval (approximating  $6/5$ ) six degrees ( $\sim 327c$ ) above the major third. An interval of  $22 - 13 = 9$  degrees, approximating a  $4/3$  perfect fourth separates the fifth (approximating the  $3/2$  perfect fifth resulting from the addition of

<sup>1</sup>To calculate the size in cents of an interval of an ET, simply divide the number of degrees by the size of the ET, and then multiple by 1200c

the major and minor third intervals) from the bottom note in the next octave. So we have intervals of 7, 6 and 9 degrees between the notes (in ascending order). Since the largest of these is 9 degrees, we set the outer interval at the remainder – 13 degrees – to minimize it. The classic major triad approximating 4:5:6, and its inversions, tuned to 22-TET, then, is labelled [0,7,13].

Figure 4.2 displays the average perceived stability of each triad type. Error bars are from 95% confidence intervals obtained from 1000 bootstrapped samples.

We are not surprised to see [0,7,13] received the highest ratings, considering it approximates the most consonant triad of 12-TET. The next most stable triad is [0,8,13]. This is 22-TET's next closest approximation to the major triad of 12-TET, approximating the frequency ratios 14:18:21 more closely than 4:5:6. To differentiate the two triads, we label [0,7,13] the *classic* major triad of 22-TET, and [0,8,13] the *supermajor* triad. This naming scheme follows from the names for the alternative major thirds – those approximating 5/4 (classic major) or 9/7 (supermajor) – from the interval naming scheme of the author's design (Hearne, 2020b). Minor triads are labelled similarly: Where the interval 7/6 is often referred to as a *subminor* third, in contrast to the more 'classic' minor third 6/5 (Hearne, 2020a). Subminor and classic minor triads are defined following this, similarly to supermajor and classic major triads. Five more triads stand out as being distinctly more stable than the others – [0,6,13], [0,4,13], [0,7,12], [0,4,9], and [0,4,11] – in order of highest to lowest average stability. [0,6,13] is 22-TET's classic minor triad, approximating the frequency ratios 10:12:15. [0,4,13] represents 22-TET's best approximations of the suspended 2nd or suspended 4th triad (equivalent by inversion), which may be written at arbitrary absolute pitch as G-A-D, A-D-G or D-G-A. We'll call the inversional equivalence class of the triad the suspended triad. [0,7,12] perhaps sounds to participants as a major triad, though one that has been "squished", or is simply mistaken for or reminiscent of a major triad, as, perhaps, is the supermajor [0,8,13]. [0,4,9] represents a subminor seventh without a fifth, approximating 12:14:21. Finally, [0,4,11] represents a dominant seventh without a fifth, approximating 4:5:7.

### 4.6.2 Hypothesis test

A LOOIC comparison reveals that removing Triad from a representative model significantly weakens it, with a difference of 835.4, for a SE of 67.0. Our hypothesis (H1) is confirmed: Triad types possess different intrinsic stabilities.

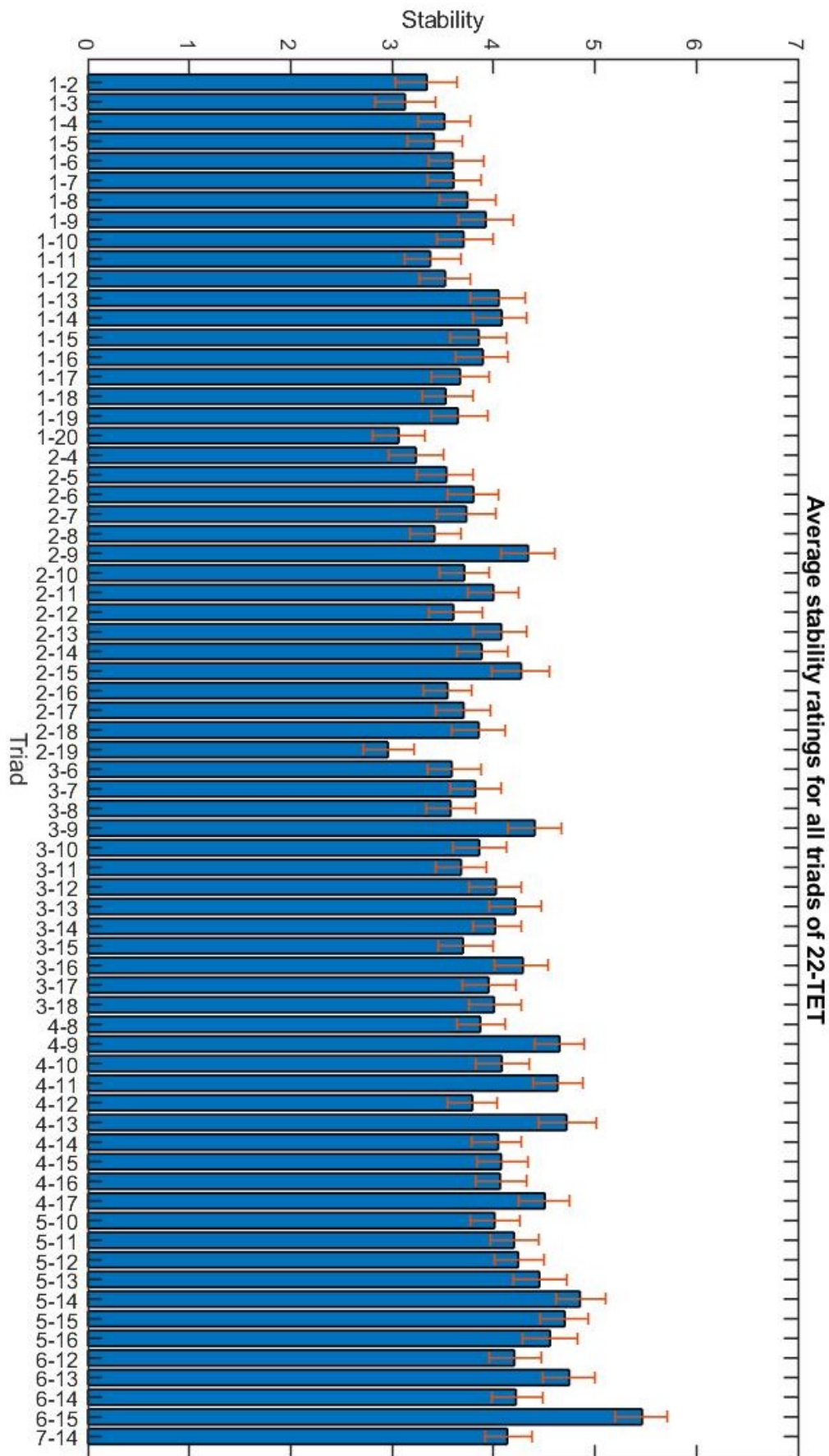


FIGURE 4.2: Average stability ratings for all triads of 22-TET

### 4.6.3 Bayesian ordinal mixed effects regression model

Through a mixed effects model we can control for our other effects in our assessment of perceived stability of triads. The descriptive model, summarized in Figure 4.2 is helpful to see, but we will look to this mixed effects model for our stability values for triads. Immediately obvious in the results of our Bayesian model was the fact that one triad was a clear outlier – [0,7,13] unsurprisingly. Controlling for the other variables in the model, the perceived stability was much higher for this triad than for any other. Considering that the dummy (reference) triad in Experiments 1 and 2 was the major triad (and its inversions), we re-ran the model with this triad as the reference triad against which the difference in stability of the others are measured.

Table 4.1 displays a selection of effects from the model after it is re-run (Model 4). *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution. Intercepts represent ‘cutpoints’ between ordinal values of ratings and the Intercept values are of a latent continuous variable; the Estimate value corresponds to the change in this latent variable that is associated with an increase of 2 standard deviations in the value of the effect for continuous variables, or with an increase from a value of 0 to a value of 1 for binary variables.

Ignoring categorical effects, the table displays intercepts, statistically significant effects, and their conditional main effects. Comparisons of each triad with the major triad are shown, as well as the interaction with Triad and MusSoph, for each comparison with the major triad, as many of these are significant. Additionally, any other interaction with Triad that is significant for comparison with the major triad is shown. The reader may notice that [0,7,13] is missing. [0,7,13] is the classic major triad, against which all triads are compared.

TABLE 4.1: Model 4 selected population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-5.11	0.23	-5.56	-4.66	733	1.00
Intercept[2]	-3.48	0.22	-3.92	-3.04	729	1.00
Intercept[3]	-2.26	0.22	-2.69	-1.82	733	1.00
Intercept[4]	-1.46	0.22	-1.89	-1.02	737	1.00
Intercept[5]	-0.14	0.22	-0.57	0.30	750	1.00
Intercept[6]	1.33	0.22	0.90	1.77	790	1.00
MusSoph	0.24	0.34	-0.43	0.91	892	1.01
Height	-0.16	0.22	-0.59	0.27	991	1.01
RelHeight	-0.06	0.07	-0.21	0.08	5710	1.00
Height <sup>2</sup>	-0.74	0.23	-1.20	-0.30	2954	1.00
RelHeight <sup>2</sup>	-0.39	0.13	-0.66	-0.13	3557	1.00
Previous	0.41	0.13	0.16	0.67	2112	1.00
Count	-0.76	0.29	-1.33	-0.20	3307	1.00
TrialNo	0.54	0.37	-0.19	1.28	1427	1.01
TrialNo <sup>2</sup>	-0.19	0.70	-1.57	1.20	2071	1.00
[0,1,2]	-2.67	0.33	-3.32	-2.02	1520	1.00
[0,1,3]	-2.66	0.31	-3.27	-2.07	1360	1.00
[0,1,4]	-1.82	0.29	-2.38	-1.25	1223	1.00
[0,1,5]	-2.13	0.30	-2.70	-1.55	1265	1.00
[0,1,6]	-1.72	0.28	-2.26	-1.18	1125	1.00
[0,1,7]	-2.16	0.28	-2.71	-1.62	1092	1.00
[0,1,8]	-1.69	0.27	-2.23	-1.16	1022	1.00
[0,1,9]	-1.87	0.28	-2.41	-1.33	991	1.00
[0,1,10]	-1.62	0.31	-2.22	-1.00	1184	1.00
[0,1,11]	-2.19	0.28	-2.73	-1.65	1079	1.00
[0,2,3]	-2.61	0.33	-3.28	-1.97	1530	1.00
[0,2,4]	-2.37	0.28	-2.91	-1.82	1149	1.00
[0,2,5]	-1.92	0.31	-2.54	-1.31	1306	1.00
[0,2,6]	-1.71	0.28	-2.26	-1.17	1105	1.00
[0,2,7]	-2.02	0.28	-2.58	-1.46	1092	1.00
[0,2,8]	-2.42	0.29	-3.00	-1.84	1250	1.00
[0,2,9]	-0.78	0.30	-1.36	-0.20	1205	1.00
[0,2,10]	-1.92	0.27	-2.46	-1.39	1156	1.00
[0,2,11]	-1.76	0.28	-2.30	-1.22	946	1.00
[0,2,12]	-2.10	0.28	-2.64	-1.55	1068	1.00
[0,3,4]	-1.89	0.30	-2.46	-1.31	1324	1.00
[0,3,5]	-2.81	0.29	-3.38	-2.26	1010	1.00
[0,3,6]	-2.01	0.28	-2.57	-1.47	1193	1.00
[0,3,7]	-1.85	0.28	-2.41	-1.29	1249	1.00
[0,3,8]	-2.05	0.27	-2.57	-1.52	1017	1.00
[0,3,9]	-1.09	0.32	-1.74	-0.48	1262	1.00
[0,3,10]	-1.55	0.27	-2.06	-1.03	980	1.00
[0,3,11]	-2.05	0.27	-2.59	-1.52	955	1.00
[0,3,12]	-1.31	0.30	-1.91	-0.73	1036	1.00
[0,4,5]	-2.17	0.28	-2.71	-1.61	1085	1.00
[0,4,6]	-1.43	0.27	-1.97	-0.90	1013	1.00
[0,4,7]	-0.99	0.29	-1.57	-0.41	1292	1.00
[0,4,8]	-1.58	0.27	-2.10	-1.05	1091	1.00
[0,4,9]	-0.71	0.30	-1.29	-0.13	1160	1.00
[0,4,10]	-1.44	0.30	-2.03	-0.86	1163	1.00
[0,4,11]	-0.96	0.33	-1.61	-0.32	1339	1.00

*Continued on next page*

Table 4.1 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
[0,4,12]	-1.52	0.27	-2.04	-0.99	1051	1.00
[0,4,13]	-0.34	0.33	-0.99	0.30	1485	1.00
[0,5,6]	-1.96	0.28	-2.53	-1.41	1009	1.00
[0,5,7]	-1.74	0.27	-2.27	-1.21	1014	1.00
[0,5,8]	-1.51	0.26	-2.03	-1.00	1087	1.00
[0,5,9]	-0.53	0.30	-1.12	0.06	1171	1.00
[0,5,10]	-1.59	0.26	-2.11	-1.07	859	1.00
[0,5,11]	-1.18	0.28	-1.74	-0.64	1096	1.00
[0,5,12]	-1.30	0.26	-1.81	-0.78	978	1.00
[0,5,13]	-1.03	0.33	-1.66	-0.38	1507	1.00
[0,6,7]	-1.81	0.31	-2.44	-1.20	1390	1.00
[0,6,8]	-1.97	0.27	-2.49	-1.44	1091	1.00
[0,6,9]	-1.29	0.26	-1.79	-0.78	938	1.00
[0,6,10]	-1.35	0.27	-1.88	-0.82	1032	1.00
[0,6,11]	-0.86	0.30	-1.46	-0.27	1274	1.00
[0,6,12]	-1.36	0.28	-1.92	-0.80	1180	1.00
[0,6,13]	-0.53	0.34	-1.18	0.16	1326	1.00
[0,6,14]	-1.27	0.28	-1.81	-0.72	1017	1.00
[0,7,8]	-1.71	0.30	-2.30	-1.12	1248	1.00
[0,7,9]	-1.10	0.30	-1.69	-0.52	1055	1.00
[0,7,10]	-1.56	0.30	-2.13	-0.98	1279	1.00
[0,7,11]	-1.06	0.28	-1.61	-0.51	1080	1.00
[0,7,12]	-0.84	0.34	-1.51	-0.19	1909	1.00
[0,7,14]	-1.08	0.27	-1.61	-0.56	1049	1.00
[0,8,9]	-1.14	0.26	-1.66	-0.62	970	1.00
[0,8,10]	-1.74	0.28	-2.29	-1.18	1138	1.00
[0,8,11]	-1.42	0.26	-1.93	-0.91	1031	1.00
[0,8,12]	-1.54	0.28	-2.09	-1.00	997	1.00
[0,8,13]	-0.40	0.27	-0.92	0.15	1070	1.00
[0,9,10]	-1.30	0.29	-1.87	-0.72	1317	1.00
[0,9,11]	-1.18	0.27	-1.72	-0.66	1051	1.00
[0,9,12]	-1.32	0.28	-1.86	-0.77	949	1.00
[0,10,11]	-1.91	0.28	-2.45	-1.36	1008	1.00
MusSoph:[0,1,2]	-2.22	0.52	-3.25	-1.21	2061	1.00
MusSoph:[0,1,3]	-2.20	0.47	-3.14	-1.29	1670	1.00
MusSoph:[0,1,4]	-2.10	0.42	-2.93	-1.28	1594	1.00
MusSoph:[0,1,5]	-1.46	0.41	-2.27	-0.67	1550	1.00
MusSoph:[0,1,6]	-0.89	0.39	-1.65	-0.14	1450	1.00
MusSoph:[0,1,7]	-0.99	0.44	-1.83	-0.13	1749	1.00
MusSoph:[0,1,8]	-1.47	0.47	-2.39	-0.53	1822	1.00
MusSoph:[0,1,9]	-0.48	0.39	-1.26	0.29	1535	1.00
MusSoph:[0,1,10]	-1.44	0.44	-2.30	-0.57	1853	1.00
MusSoph:[0,1,11]	-1.00	0.41	-1.82	-0.20	1733	1.00
MusSoph:[0,2,3]	-1.45	0.53	-2.50	-0.39	1986	1.00
MusSoph:[0,2,4]	-1.79	0.43	-2.64	-0.95	1778	1.00
MusSoph:[0,2,5]	-1.25	0.51	-2.27	-0.25	1910	1.00
MusSoph:[0,2,6]	-1.00	0.40	-1.79	-0.22	1381	1.00
MusSoph:[0,2,7]	-1.26	0.41	-2.06	-0.46	1572	1.00
MusSoph:[0,2,8]	-0.75	0.42	-1.57	0.06	1730	1.00
MusSoph:[0,2,9]	-1.11	0.43	-1.93	-0.28	1842	1.00
MusSoph:[0,2,10]	-1.16	0.42	-1.97	-0.35	1723	1.00
MusSoph:[0,2,11]	-1.12	0.39	-1.90	-0.35	1500	1.00

Continued on next page

Table 4.1 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
MusSoph:[0,2,12]	-1.41	0.38	-2.17	-0.67	1423	1.00
MusSoph:[0,3,4]	-1.59	0.45	-2.47	-0.70	1907	1.00
MusSoph:[0,3,5]	-1.90	0.43	-2.76	-1.06	1543	1.00
MusSoph:[0,3,6]	-1.03	0.41	-1.84	-0.23	1637	1.00
MusSoph:[0,3,7]	-0.43	0.43	-1.27	0.41	1545	1.00
MusSoph:[0,3,8]	-1.06	0.40	-1.84	-0.29	1561	1.00
MusSoph:[0,3,9]	0.28	0.46	-0.62	1.18	1729	1.00
MusSoph:[0,3,10]	-1.10	0.40	-1.89	-0.30	1598	1.00
MusSoph:[0,3,11]	-0.86	0.38	-1.62	-0.11	1268	1.00
MusSoph:[0,3,12]	-0.57	0.47	-1.50	0.34	1963	1.00
MusSoph:[0,4,5]	-1.60	0.40	-2.40	-0.81	1683	1.00
MusSoph:[0,4,6]	-0.79	0.39	-1.56	-0.02	1534	1.00
MusSoph:[0,4,7]	-0.83	0.45	-1.70	0.04	1946	1.00
MusSoph:[0,4,8]	-0.56	0.38	-1.31	0.18	1293	1.00
MusSoph:[0,4,9]	-0.10	0.47	-1.02	0.82	1916	1.00
MusSoph:[0,4,10]	-0.40	0.48	-1.34	0.54	1923	1.00
MusSoph:[0,4,11]	-0.44	0.51	-1.45	0.55	1729	1.00
MusSoph:[0,4,12]	-0.79	0.37	-1.51	-0.06	1409	1.00
MusSoph:[0,4,13]	-0.52	0.50	-1.50	0.45	1973	1.00
MusSoph:[0,5,6]	-1.83	0.40	-2.61	-1.05	1588	1.00
MusSoph:[0,5,7]	-1.14	0.38	-1.89	-0.42	1504	1.00
MusSoph:[0,5,8]	-0.55	0.39	-1.30	0.20	1499	1.00
MusSoph:[0,5,9]	-0.01	0.47	-0.94	0.91	1954	1.00
MusSoph:[0,5,10]	-0.67	0.37	-1.40	0.03	1296	1.00
MusSoph:[0,5,11]	-1.13	0.41	-1.93	-0.33	1551	1.00
MusSoph:[0,5,12]	-0.45	0.39	-1.23	0.31	1607	1.00
MusSoph:[0,5,13]	-0.34	0.52	-1.35	0.70	2122	1.00
MusSoph:[0,6,7]	-1.14	0.50	-2.10	-0.17	1920	1.00
MusSoph:[0,6,8]	-0.82	0.37	-1.54	-0.11	1307	1.00
MusSoph:[0,6,9]	0.08	0.36	-0.65	0.78	1310	1.00
MusSoph:[0,6,10]	-0.56	0.38	-1.31	0.19	1370	1.00
MusSoph:[0,6,11]	-0.34	0.47	-1.26	0.59	1792	1.00
MusSoph:[0,6,12]	-1.45	0.42	-2.28	-0.62	1646	1.00
MusSoph:[0,6,13]	0.10	0.54	-0.96	1.17	1959	1.00
MusSoph:[0,6,14]	-0.88	0.41	-1.69	-0.08	1461	1.00
MusSoph:[0,7,8]	-1.27	0.44	-2.13	-0.41	1682	1.00
MusSoph:[0,7,9]	-0.48	0.44	-1.38	0.38	1705	1.01
MusSoph:[0,7,10]	-1.09	0.41	-1.89	-0.28	1675	1.00
MusSoph:[0,7,11]	-0.97	0.43	-1.80	-0.12	1832	1.00
MusSoph:[0,7,12]	0.03	0.54	-1.01	1.10	2448	1.00
MusSoph:[0,7,14]	-0.27	0.38	-1.03	0.47	1463	1.00
MusSoph:[0,8,9]	-1.23	0.39	-1.99	-0.46	1596	1.00
MusSoph:[0,8,10]	-1.25	0.38	-2.00	-0.52	1546	1.00
MusSoph:[0,8,11]	-0.92	0.39	-1.70	-0.15	1500	1.00
MusSoph:[0,8,12]	-0.84	0.41	-1.64	-0.05	1610	1.00
MusSoph:[0,8,13]	-0.05	0.38	-0.80	0.69	1404	1.00
MusSoph:[0,9,10]	-0.65	0.43	-1.48	0.19	1674	1.00
MusSoph:[0,9,11]	-0.77	0.38	-1.52	-0.03	1436	1.00
MusSoph:[0,9,12]	-0.52	0.38	-1.28	0.21	1350	1.00
MusSoph:[0,10,11]	-0.94	0.42	-1.77	-0.11	1731	1.00
TrialNo:[0,1,8]	-1.10	0.44	-1.96	-0.25	1695	1.00
TrialNo <sup>2</sup> :[0,7,14]	-1.61	0.72	-3.03	-0.18	4601	1.00



Given that stability ratings of chords reflected their consonance in Experiments 1 and 2, we hypothesize that we may find the same for microtonal triads. Indeed, our second hypothesis, which is not tested here, tests to see if stability ratings of triads can indeed be modelled by consonance – represented by an additive model of harmonicity, spectral entropy, harmonic entropy, familiarity and sensory dissonance. We are not surprised to see then that all triads including intervals smaller than 4 degrees ( $\sim 218c$ ) are given significantly lower stability ratings than the classic major triad, considering the relatively infrequent use of such triads in Western tonal-harmonic music, and our understanding of them as dissonant. Only a small number of triads were not rated significantly lower than the classic major triad, namely, Triads [0,4,13], [0,6,13], [0,8,13], and [0,5,9]. Table 4.2 lists the 16 most stable triads, along with their name and the size in cents and frequency ratios approximated by the inversion the name describes.

TABLE 4.2: Perceived stability of the 16 most stable triads

Rank	Triad	Name	Size in cents	Ratios approximated	Stability
1	[0,7,13]	Classic major triad	382-709	5/4, 6/5, 4/3, 4:5:6	0
2	[0,4,13]	Suspended triad	491-709	4/3, 8/7, 4/3, 6:8:9	-0.34
3	[0,8,13]	Supermajor triad	436-709	9/7, 7/6, 4/3, 14:18:21	-0.40
4	[0,6,13]	Classic minor triad	327-709	6/5, 5/4, 4/3, 10:12:15	-0.53
5	[0,5,9]	Subminor seventh (no third)	709-982	3/2, 7/6, 8/7, 4:6:7	-0.53
6	[0,4,9]	Subminor seventh (no fifth)	273-982	7/6, 3/2, 8/7, 12:14:21	-0.71
7	[0,2,9]	Classic major seventh (no fifth)	382-1082	5/4, 3/2, 15/14, 8:10:15	-0.78
8	[0,7,12]	“Squished” major triad	382-655	5/4, 7/6, 11/8, 24:30:35	-0.84
9	[0,6,11]	Harmonic diminished triad	327-600	6/5, 7/6, 7/5, 5:6:7	-0.86
10	[0,4,11]	Harmonic dominant seventh (no fifth)	382-982	5/4, 7/5, 8/7, 4:5:7	-0.96
11	[0,4,7]	Classic major add 9 (no fifth)	218-382	8/7, 10/9, 8/5, 8:9:10	-0.99
12	[0,5,13]	Subminor triad	273-709	7/6, 9/7, 4/3, 6:7:9	-1.03
13	[0,7,11]	Classic major flat 5	382-600	5/4, 8/7, 7/5, 12:15:17	-1.06
14	[0,7,14]	Classic augmented triad	382-765	5/4, 5/4, 9/7, 16:20:25	-1.08
15	[0,3,9]	Classic minor seventh (no fifth)	327-1036	6/5, 3/2, 10/9, 5:6:9	-1.09
16	[0,7,9]	Classic major seventh (no third)	709-1082	3/2, 5/4, 15/14, 8:12:15	-1.10

The size in cents (rounded to the nearest cent) and the simplest possible frequency ratios given the treatment of 22-TET as a temperament with consistent mappings of harmonics up to 17, excluding 13, to its intervals (“22edo”, 2020) is listed for each of the intervals, in order, and of the triad as a whole, for the root position of the triad given its name. 95% credibility intervals (CIs) on the stability values can be found in Table 4.1 (apart from for [0,7,13], as it is the dummy coded triad, which does not have 95% CIs).

Figure 4.3 notates the 16 triads detailed in Table 4.2 using Ups and Downs Notation (Giedraitis, n.d.), a system developed by Kite Giedraitis to notate almost

## 16 most stable triads of 22-TET

The figure displays 16 triads in 22-TET, arranged in four rows. Each triad is shown on a treble clef staff with a key signature of one sharp (F#). The notes are represented by circles with accidentals (sharps, flats, naturals) and accidentals (ups and downs) indicating their position relative to the diatonic scale. Below each triad is its interval set and a descriptive name.

Triad	Interval Set	Name
1.	[0,7,13]	classic major
2.	[0,4,13]	suspended
3.	[0,8,13]	supermajor
4.	[0,6,13]	classic minor
5.	[0,5,9]	subminor seventh (no third)
6.	[0,4,9]	subminor seventh (no fifth)
7.	[0,2,9]	classic major seventh (no fifth)
8.	[0,7,12]	"squished" major
9.	[0,6,11]	harmonic diminished
10.	[0,4,11]	harmonic dominant seventh (no fifth)
11.	[0,4,7]	classic major add 9 (no fifth)
12.	[0,5,13]	subminor
13.	[0,7,11]	classic major flat 5
14.	[0,7,14]	classic augmented
15.	[0,3,9]	classic minor seventh (no fifth)
16.	[0,7,9]	classic major seventh (no third)

FIGURE 4.3: The 15 most stable triads of 22-TET, notated using Ups and Downs

all ETs up to 72. Only two new accidentals are introduced: an *up* – ‘ $\wedge$ ’ – and a *down* – ‘ $\vee$ ’ – which signify raising or lowering a note respectively by a single degree of the ET. Standard diatonic notation in 22-TET, to which ups and downs are added, uses the nominals on the staff and the sharps and flats determined by the circle of fifths using the ET’s best fifth (the ETs closest approximation to  $3/2$ ). Figure 4.4, printed as Figure 4.1 in Giedraitis (n.d.), downloaded from the page for 22-TET on the Xenharmonic Wiki “22edo” (2020) provides a guide for the notation of 22-TET in Ups and Downs. Audio examples of these triads can be found at [https://en.xen.wiki/w/The\\_16\\_most\\_stable\\_triads\\_of\\_22edo](https://en.xen.wiki/w/The_16_most_stable_triads_of_22edo) (*edo* is short for *equal divisions of the octave*, and so 22edo is equivalent to 22-TET as far as we need to be concerned in this thesis).

[0,4,13] is the suspended triad. The triad [0,5,9] represents a subminor seventh



FIGURE 4.4: 22-TET Ups and Downs notation guide, from *Tibia 22edo ups and downs guide 1* by Kite Giedraitis, 2016. Downloaded from [https://en.xen.wiki/w/File:Tibia\\_22edo\\_ups\\_and\\_downs\\_guide\\_1.png](https://en.xen.wiki/w/File:Tibia_22edo_ups_and_downs_guide_1.png).

Used with permission.

tetrad without a third, which may be written at arbitrary absolute pitch as D-A-C (or A-C-D or C-D-A).  $[0,8,13]$  is 22-TET's *supermajor* triad, approximating the frequency ratios 14:18:21.  $[0,6,13]$  is 22-TET's classic minor triad. We were surprised to see 22-TET's *subminor* triad (approximating 6:7:9) rated less stable (after controlling for other effects) than all of these, considering that its representation as frequency ratios is so simple, but we understand this is only a clue to consonance. It might be the similarity of the supermajor triad to the 12-TET major triad, which received high stability ratings, that led to its high ratings, despite its more complex representation as approximated frequency ratios.

Considering the interaction of Triad with MusSoph, for all triads apart from  $[0,1,9]$ ,  $[0,9,10]$ ,  $[0,2,8]$ ,  $[0,7,9]$ ,  $[0,3,7]$ ,  $[0,3,9]$ ,  $[0,3,12]$ ,  $[0,9,12]$ ,  $[0,6,9]$ ,  $[0,5,8]$ ,  $[0,4,7]$ ,  $[0,4,8]$ ,  $[0,4,9]$ ,  $[0,4,10]$ ,  $[0,4,11]$ ,  $[0,4,13]$ ,  $[0,6,10]$ ,  $[0,5,9]$ ,  $[0,5,10]$ ,  $[0,5,12]$ ,  $[0,5,13]$ ,  $[0,8,13]$ ,  $[0,7,12]$ ,  $[0,6,11]$ ,  $[0,6,13]$  and  $[0,7,14]$  the positive influence of higher musical sophistication on ratings is significantly lower (noting the positive, though insignificant effect of MusSoph on ratings for the classic major triad). These include the Triads  $[0,4,13]$ ,  $[0,5,9]$ ,  $[0,5,13]$  and  $[0,6,13]$  from immediately above, along with many more. The first two triads –  $[0,1,9]$ , and  $[0,9,10]$  – are 22-TET's perfect 4th and perfect 5th respectively – the most consonant intervals within an octave – with a note added a single degree above the lower note (or perfect fourths with a note added a single degree above the lower note or the upper note respectively). We assume that if intrinsic stability can be modelled by consonance – an additive model of harmonicity, spectral entropy, harmonic entropy, familiarity and sensory dissonance – as will be tested analysis of our second hypothesis in the future, the high consonance (considering the high stability) of these triads is due in part to the high consonance of those intervals. It is not immediately clear why  $[0,2,8]$ ,  $[0,3,7]$ ,  $[0,5,8]$ ,  $[0,6,10]$  and  $[0,5,10]$  might be considered consonant. We note that for  $[0,2,8]$ ,  $[0,5,8]$

and [0,5,10] at least, positive influence of MusSoph on ratings is lower, if not quite significantly so. [0,7,9] also includes a perfect fifth, this time with a note added two degrees below the lower note. [0,3,9] represents a classic minor seventh with no fifth, approximating the frequency ratio 5:6:9; [0,3,12] approximates 11:12:16, or in another inversion 6:8:11; [0,9,12] approximates 10:11:15. These last two triads can be thought of as suspended triads that have had one note altered by a quarter-tone. [0,6,9] represents a classic minor seventh without a third, approximating 10:15:18; [0,4,8] approximates 7:8:9; [0,4,7] approximates 8:9:10, a classic major add 9 tetrad without the fifth; [0,4,9] represents a subminor seventh without a fifth, approximating 12:14:21; [0,4,10] approximates 8:9:11; [0,4,11] represents a dominant seventh without a fifth, approximating 4:5:7. Since the whole tetrad would approximate the harmonic series segment 4:5:6:7, we call this tuning of the dominant seventh tetrad the *harmonic dominant seventh*; [0,7,12] is our “squished” major triad, and [0,5,12] perhaps sounds to participants similarly as a “squished” minor triad; [0,5,13] is the subminor triad approximating 6:7:9, or perhaps they are simply close enough approximations to the major and minor triads to function as them perceptually; [0,6,11] represents the harmonic diminished triad, approximating 5:6:7; [0,7,14] is the classic augmented triad approximating 16:20:25.

From Table 4.2 we can see the influence in the ratings of the simplicity of frequency ratios approximated by both the constituent intervals and of the triad as a whole, as well the pull of the familiarity of the 12-TET major triad. Analysis of the second hypothesis in further research will assess this observation.

It is perhaps interesting to note that 9 of the most stable 11 triads can be found in the following three tetrads: Classic major seventh (classic major triad, classic minor triad and classic major seventh no fifth), subminor seventh (supermajor triad, subminor seventh no third, subminor seventh no fifth and subminor triad), and harmonic dominant seventh (harmonic diminished triad, harmonic dominant seventh no fifth).

The least stable 5 triads, ordered from least to most stable, are [0,3,5], [0,1,2], [0,1,3] and [0,2,3], and [0,2,4]. We are not surprised that these triads all contain two intervals smaller than 200c.

Previous, Count and TrialNo are significant and positive, suggesting that ratings are positively influenced by the rating given to the previous trial, and that ratings are higher for later trials, and for probes for which the constituent pitch classes have been heard more in the experiment up until the time of the trial.

A significant negative interaction exists between [0,1,8] and TrialNo, suggesting that [0,1,8] received lower stability ratings (as compared to the classic major triad) later in the experiment.

Finally, a significant negative interaction exists between [0,7,14] – the Augmented triad – and TrialNo<sup>2</sup>, suggesting that [0,7,14] was rated lower (as compared to the classic major triad) towards both the beginning and end of the experiment.

We don't have an explanation at this stage for the significance of the last two effects.

## 4.7 Discussion / Conclusion

Unsurprisingly, 22-TET's classic major triad – [0,7,13] – emerges as the most stable triad. From an observation of Figure 4.3, [0,8,13], 22-TET's supermajor triad looks to be the next most stable, followed by Triads [0,6,13] [0,4,13], [0,7,12], [0,4,9], and [0,4,11]. A Bayesian mixed effects model is run, controlling for many contextual effects. [0,7,13] was found to be more stable than all triads apart from [0,4,13], [0,5,9], [0,8,13] and [0,6,13]. The 16 most stable triads from this model include the 7 triads found to be the most stable in the descriptive analysis (16 and 7 seemed like appropriate stopping points in both cases as there was a comparatively large drop to the stability of the 17th and 8th most stable triads respectively). The mixed effects model was compared via LOOIC to a model with the effect of Triad, and found to significantly outperform it, confirming our first hypothesis that triads, at least in 22-TET, possess differing intrinsic stabilities.

Instead of using categorical effect of Triad, a future analysis will assess the intrinsic stability of the triads with an additive model of sensory dissonance, harmonicity, spectral entropy, harmonic entropy and familiarity. The high stability ratings given to the less accurate approximations to a 12-TET major triad, i.e., [0,8,13], the triad approximating 14:18:21 triad and the "Squished" major triad – [0,7,12] – are interesting, suggesting the major triad has a stronger "gravity" or "pull" than anticipated. It will be interesting to see if our additive model concerning our second hypothesis is able to account for it. We should consider including effects of the distance of each triad not just from the closest 12-TET triad, but to the 12-TET major triad to account for familiarity in our model testing our second hypothesis.

The average stability ratings for triads in this experiment are used in a distributional analysis in the next chapter, which leads to the selection of seven-note scales of 22-TET to be tested in Experiment 4 and 5. After some terms are defined it begins with a review of the literature of scale features that might affect how well a scale might support harmonic tonality. New features are defined, and the max, median and min stability of the tertian triads available in the scales are included. A cluster analysis of the values of a reduced set of features of all seven-note scales of 22-TET leads to set of scales for testing in a pair of experiments detailed in Chapter 6.

## Chapter 5

# Distributional Analysis of $n$ -dimensional Feature Space for 7-note Scales in 22-TET

### 5.1 Introduction

This chapter, adapted from Hearne et al. (2019), details the selection of a small set of scales representative of all 7752 7-note scales of 22-TET, in terms of features that may affect the possible suitability of these scales for harmonic tonality. We note that the relative influences of the features of scales in 12-TET have been difficult to disentangle: The diatonic scale performs highly according to almost all measures. In 22-TET however, the features are spread differently across different scales. We sought here to establish a set of 7-note scales in 22-TET that exemplify the major clusters within the whole population of scales. After a review of the literature on scale features we select those scale features that have potential relevance to harmonic tonality and calculate their values for every 7-note scale in 22-TET. This feature space is then reduced by the step-by-step removal of features whose values may be most fully expressed as linear combinations of the others. A  $K$ -medoids cluster analysis leads finally to the selection of 11 exemplar scales, which include approximations of four different tunings of the diatonic scale in just intonation. This exemplar set is to be used in Experiments 4 and 5, detailed in Chapter 6 (as discussed at the chapter's conclusion, Section 5.5, a slightly different set of 8 exemplar scales were used in Experiments 4 and 5). Before we can discuss our features or the analysis any further, we must first define the terms we will be using in this chapter.

## 5.2 Definitions

1. *Degree*: The smallest interval of an equal temperament (ET). For example, one degree of 12-TET is a semitone of 100c.
2. *Just Intonation (JI)*: A tuning system in which all intervals are represented as, and tuned to, integer ratios of frequencies of vibration.
3. *Specific Interval*: The size of a musical interval in degrees of an ET, or as ratio of frequencies of vibration of the pitches. Also referred to as the *specific size* of a musical interval.
4. *Scale*: An equivalence class by rotation of ordered sets of intervals called steps.
5. *Step*: An interval between adjacent pitches of a scale.
6. *Mode*: A specific rotation of a scale. For example, 2212221 is the 'major' or 'Ionian' mode of the diatonic scale, whereas 2122212 is called 'Dorian mode'. In Sections 5.4 and 5.5 scales in ETs are written in their "brightest" mode (the mode in which the larger steps are most concentrated towards the beginning), unless otherwise indicated, with step sizes written in degrees of the ET. For example, the diatonic scale in 12-TET is represented as 2221221, which is 'Lydian mode'. Scales in JI are represented with the frequency ratios of its intervals above the tonic, in the most appropriate mode.
7. *Generic Interval*: When it exists in a scale, a musical interval has both a specific and a generic size. The *generic interval class* of a musical interval, or equivalently, a generic interval, is the number of steps of a scale subtended by a musical interval, plus one. In the diatonic scale each generic interval comes in two specific sizes. For the generic intervals of a second, third, sixth and seventh, the smaller of these is labelled 'minor' and the larger is labelled 'major'.
8. *Perfect Fifth*: There are multiple ways to define a perfect fifth. within the context of a diatonic scale, 'fifth' indicates the generic interval, i.e., that it is subtended by 4 steps of the diatonic scale, and then 'perfect' indicates the specific size of this fifth – which in 12-TET is 7 degrees or equivalently, 7 semitones – in the same way that 'major' or 'minor' do for seconds, thirds, sixths and sevenths. The *JI perfect fifth*, however, is defined as the specific interval represented by the frequency ratio '3/2'. In this paper, 'perfect fifth' refers to the



JI perfect fifth, and its approximation as either 7 degrees of 12-TET, or as 13 degrees of 22-TET, and not necessarily to a musical interval subtended by 4 steps of the diatonic scale.

9. *Period*: The interval at which a scale repeats. Typically, an octave (an interval ratio of 2/1).
10. *Generator*: A specific interval which, when stacked and after the resulting pitches are transposed by octave to within a single period, produces a *generated scale* or a tuning system.
11. *Meantone*: The tuning system generated by a flattened perfect fifth, with the period of an octave. Six meantone generators stacked above a note generates the meantone tuning of the diatonic scale, in Lydian mode.
12. *Tetrachord*: A segment of 4 adjacent pitches of a scale. Tetrachords comprising identical intervals in an identical order are said to be identical, regardless of their pitch height. Tetrachord is defined here only for 7-note scales. For example, the tetrachord 221 occurs twice in the Lydian mode of the diatonic scale – 2221221. Tetrachords often span an interval that approximates the frequency ratio 4/3, but they do not have to.

## 5.3 Review

The scale features we consider may be divided into six groups which will be defined below: Generator complexity,  $R$ -ad entropy, redundancy, coherence and evenness, consonance, and tetrachordality. The diatonic scale in 12-TET boasts equal lowest generator complexity and  $R$ -ad entropy and equal highest redundancy for 7-note scales in 12-TET. It is also the maximally even 7-note scale in 12-TET, and 12-TET's only omnitetra-chordal scale. The diatonic scale also maximizes the number of constituent consonant triads<sup>1</sup>. Since in 12-TET the diatonic scale holds what is almost a monopoly on many of these features, we need to look elsewhere if we are to tease them apart. 22-TET is chosen as it is the smallest tuning wherein a single scale no longer stands out as such an outlier in terms of its values for these features, but where all the features we define exist across an appropriate range of values in some scales. We choose also to limit our analysis to scales of 7 notes to reduce the size

of our set and simplify our analysis. Scale features described in the literature are reviewed and some new features are defined. We begin with redundancy.

### 5.3.1 Redundancy

After commenting on the many different conceptions of a musical scale that exist in the literature, Carey (2002) suggests that a pitch class set can be considered a scale, ‘when its generic intervals efficiently organize and encode its specific intervals. Put simply, a scale is that kind of pitch-class set in which it makes sense to think about intervals generically’ (Carey, 2002, p. 5).

We remind the reader that a specific interval is defined as ‘the size of a musical interval in degrees of an ET, or as a ratio of frequencies of vibration of the pitches’, whereas a generic interval is defined as ‘the number of steps of a scale subtended by a musical interval, plus one’. These concepts are the most important to understand in order to follow the review of scale features. We will use the diatonic scale in 12-TET, in its most common mode ‘major’ or ‘Ionian’ – 2212221 – as an example throughout the review. We list the specific size in degrees of 12-TET or equivalently, semitones, of each generic interval – i.e., each 2nd, 3rd, 4th, 5th, 6th and 7th:

- 2nds: 2,2,2,1,2,2,1
- 3rds: 4,4,3,3,4,3,3
- 4ths: 6,5,5,5,5,5,5
- 5ths: 7,7,7,7,7,7,6
- 6ths: 9,9,8,8,9,9,8
- 7ths: 11,10,10,10,11,10,10

We can see that there is a close relationship between the specific and generic intervals of the diatonic scale. Many pitch-class sets do not possess such a relationship. Redundancy and coherence concern this relationship between specific and generic intervals. *Redundancy* concerns the certainty with which a generic interval infers a specific interval, while *coherence* (discussed in the following subsection) concerns the inverse: the certainty with which a specific interval infers a generic interval.

Considering redundancy, Rothenberg defines the *variety* of a generic interval as the number of specific sizes it comes in. For example, in the 12-TET diatonic scale – 2221221 – we can immediately see that 2nds (steps) come in 2 sizes: 1 and 2 degrees.

We can see from our examination of the scale's specific and generic intervals above that all generic intervals come in two specific sizes. The significance of this will be discussed below. *Mean variety* (Rothenberg, 1977) and *maximum variety* follow directly from this, considering all the generic intervals of the scale (up to  $N - 1$ , where  $N$  is the cardinality of the scale). For the 12-TET diatonic scale, since the variety of each generic interval is 2, the mean and maximum variety of the scale both take the value 2.

We will briefly introduce a new concept here: the generated scale. A *generated scale* is a scale that can be produced from the iterated addition of a specific interval modulo the period (Clough et al., 1999). Where a circle of fifths can produce the diatonic scale (starting on F, going up perfect fifths, we arrive at C, G, D, A, E and finally at B), we know that it is a generated scale. Wilson noted that some *generated scales* possess the property that the maximum variety is two. He calls these scales *moment of symmetry* or *MOS scales* (Wilson, 1975a, 1975b). We know now that the 12-TET diatonic scale is an MOS scale.

Clough and Douthett defined *maximally even (ME) scales* as scales in which each generic interval has either one or two adjacent specific intervals, meaning that it is 'distributed as evenly as possible' (Clough & Douthett, 1991, p. 96). ME scales are a subset of *distributionally even (DE scales)*, where each generic interval comes in either one or two specific intervals (Clough et al., 1999). The 12-TET diatonic scale, therefore, is also a DE scale, as are all MOS scales. If we look back at our list of intervals of the diatonic scale in 12-TET we can see that it is also ME: It is the ME 7-note scale of 12-TET.

Before introducing the next concept, we need to consider generated scales again briefly. A generator is, by definition, of invariant specific size. A given specific interval may be of variant generic size however, and therefore so can a generator. Scales for which a specific interval is represented by more than one generic interval class are discussed in the following subsection. A generated scale in which the generator is of invariant generic size is called a *well-formed (WF) scale* (Carey & Clampitt, 1989). WF scales come in two types: *degenerate*, the set of ETs, and *non-degenerate*, the set of scales that possess *Myhill's property* – that each generic interval comes in exactly two specific sizes (Clough & Myerson, 1985). We should be far from surprised at this point that the 12-TET diatonic scale is also (non-degenerate) WF. In non-equal scales (scales that are not ETs, but may be subsets of ETs) of prime cardinality, WF, DE and MOS are equivalent. We refer henceforth to these scales as WF.

After Myhill's property for well-formed scales, *trivalent scales* are defined such that each generic interval comes in three specific sizes. Consider the JI major scale,  $9/8, 5/4, 4/3, 3/2, 5/3, 15/8, 2/1$ . With steps of  $9/8, 10/9, 16/15, 9/8, 10/9, 9/8, 16/15$ , it is trivalent (Carey, 2007; Clampitt, 2007). Expressed as 'L', 's' and 'm' for large, small and medium steps, this may be described as LmsLmLs. In meantone temperament, the minor and major tones –  $10/9$  and  $9/8$  – are tempered to equivalence (tempering out their difference,  $81/80$ ). This leads us back to the well-formed meantone diatonic, which may be described, in the major mode, as LLsLLLs, which we have seen, tuned to 12-TET, as 2212221.

If we take any other pair of step sizes to be equivalent, we also are led to well-formed scales. i.e., taking  $10/9$  to be equivalent to  $16/15$  (tempering out  $25/24$ ) leads to LssLsLs, and taking  $9/8$  to be equal to  $16/15$  (tempering out  $135/128$ ) leads to sLssLss. This property is described by Clampitt as *pairwise well-formedness* (Clampitt, 1998).

Carey later introduces the concept of *strong  $n$ -valence* as a generalisation to a consequence of Myhill's property: 'Let  $n$  represent the number of distinct step sizes per span. If the set of  $(n)(n - 1)/2$  (positive) differences between the  $n$  step sizes is the same for each span, the set has strong  $n$ -valence' (Carey, 2007, p. 96). He conjectures that a set of odd cardinality has strong trivalence if and only if it is pairwise well-formed. This conjecture will be returned to in Section 5.4.1 below.

An instance of a pair of intervals of the same generic size which differ in specific size is called a *difference*. Carey's *sameness quotient* gives a continuous measure of the infrequency of difference in a scale, which is where a pair of intervals of the same generic size differs in specific size (Carey, 2002).

Another similar feature, which will here be called  *$n$ -chord entropy* is introduced recently in a rhythmic context by Milne and Dean (2016) and elaborated on in Milne and Herff (2020) applied to scales. Here we use it to consider the entropy of the distribution of  $n$ -chords, which are  $n$  note factors/segments of the scale (we are most familiar with  $n$ -chords when  $n$  is 4; i.e., tetrachords). The probability mass function

$$P_i(n)$$

is the number of occurrences of each different  $n$ -chord, divided by the number of notes in the scale. Then the  $n$ -chord entropy in bits is as follows:

$$E(P) = - \sum_i P_i \log_2 P_i \quad (5.1)$$

$n$ -chord entropy is defined in a scale of  $N$  notes for

$$2 \leq n \leq N - 1$$

.

### 5.3.2 Coherence and Evenness

We recall from the previous subsection that coherence concerns the certainty with which a specific interval infers a generic interval. We've already come across the existence of scales for which a specific interval does not infer a single generic interval. Such scales are said to be *improper*, where a scale is considered *proper* if no specific interval of generic interval class  $n$  is larger than any specific interval of generic interval class  $n+1$  (Rothenberg, 1975). A scale is considered to be *strictly proper* if no specific interval of generic size  $n$  is *equal to* or larger than any specific interval of generic size  $n+1$ . Strict propriety may be broken by a *contradiction*, when propriety also fails, or by an *ambiguity*, where only strict propriety fails. The diatonic scale in 12-TET is proper, but not strictly proper, with a single ambiguity – the tritone, which may be a diminished fifth or an augmented fourth.

Balzano (1982) independently introduced the concept of *coherence*, equivalent to strict propriety. He then then defined a weaker version of coherence which the diatonic scale in 12-TET passes, in which ambiguity is allowed for an interval of half an octave (the tritone).

Tuned as it was for centuries to Pythagorean intonation (a tuning system generated by pure perfect fifths – i.e., with frequency ratio exactly  $3/2$ ), the diatonic scale is improper, where the Aug 4 is larger than the dim 5. With meantone tempering it is strictly proper. Clearly a scale does not need to be strictly proper or even proper to be tonal, and accordingly we do not include binary coherence features in our analysis. Non-binary measures for coherence have also been defined, by which the various tunings of the diatonic scale receive extreme values.

Similar to his sameness quotient, in the same paper Carey introduced a coherence quotient as a continuous measure for the infrequency of failures of coherence (ambiguity or contradiction) (Carey, 2002).

Along with propriety, Rothenberg introduced *stability*, with which proper scales can be compared, defined as the portion of unambiguous intervals, out of all  $N(N - 1)$  possible intervals (Rothenberg, 1975). Unlike Carey's coherence quotient which considers both ambiguities and contradictions, Rothenburg stability concerns only ambiguities (of any degree). Given that it is only defined for proper scales we do not include it in our analysis.

Thus far no feature directly concerns the relative size of intervals in the scale. Lumma introduces two concepts intended to take this into account. The first of these – Lumma stability – is an extension of Rothenberg's stability. Lumma stability is the portion of the octave that is not covered with the spans of each generic interval class. We will look not at 12-TET initially for a worked example. The WF diatonic scale, generated by a perfect fifth, is also available in 22-TET, with a large step of 4 degrees, and a small step of 1 degree – 4414441 in the major mode – provides a more clear example. It's generic interval spans are as such:

- 2nds of 4,4,1,4,4,4,1, spanning 1-4
- 3rds of 8,5,5,8,8,5,5, spanning 5-8
- 4ths of 9,9,9,12,9,9,9, spanning 9-12
- 5ths of 13,13,10,13,13,13,13, spanning 10-13
- 6ths of 17,17,14,17,17,14,14, spanning 14-17
- 7ths of 21,18,18,21,18,18,18, spanning 18-21

0-1, 4-5, 8-9, 13-14, 17-18 and 21-22 are not covered by the spans of generic interval classes, resulting in a Lumma stability of  $6/22 = 3/11$ .

The portion of the octave more than singly covered by the spans of each generic interval class is defined as Lumma impropriety (Op de Coul, 2020). For the scale 4414441, the portion 10-12 is doubly covered (by 4ths and by 5ths), and so the Lumma impropriety is  $2/22 = 1/11$ .

We'll return to the major mode 2212221 of the diatonic scale for another example. With

- 2nds of 2,2,1,2,2,2,1, spanning 1-2

- 3rds of 4,3,3,4,4,3,3, spanning 3-4
- 4ths of 5,5,5,6,5,5,5, spanning 5-6
- 5ths of 7,7,6,7,7,7,7, spanning 6-7
- 6ths of 9,9,8,9,9,8,8, spanning 8,9
- 7ths of 11,10,10,11,10,10,10, spanning 10-11

0-1, 2-3, 4-5, 7-8, 9-10 and 11-12 are not covered, resulting in a Lumma stability of 1/2. For proper scales no portion of the scale can be more than singly covered, as generic intervals cannot cross over. Proper scales, then, including the 12-TET diatonic scale, have a Lumma impropriety of 0.

*Evenness* also directly concerns the relative sizes of intervals of a scale, measuring the similarity of the scale to an ET of the same cardinality. For more thorough definitions and formulae see (Amiot, 2009; Milne, Bulger, Herff, & Sethares, 2015). Evenness can be seen as a continuous generalization of the binary measure of maximal evenness.

### 5.3.3 *R*-ad entropy

We define *R*-ad entropy as the entropy of the distribution of subsets of *R* notes – “*R*-ads” – from a scale of cardinality *N* where *R* ranges from 2 to *N* – 1. We consider however only *R*-values of 2 and 3, corresponding to dyads and triads, as we consider larger subsets of notes to be less important to tonality. The entropy in bits is calculated using the probability mass function of the number of occurrences of each different *R*-ad, divided by the total number of *R*-ads.

### 5.3.4 Generator complexity

Generator complexity considers the compactness with which the scale can be represented in a minimum number of dimensions. Where the *Graham complexity*, after Graham Breed, is the number of generators needed to reach an interval in a scale or a 2-dimensional tuning system, we define *scalar Graham complexity (SGC)* as the minimum number of generators of a given size needed to cover a scale, across all possible sizes of generator. Let’s consider again the diatonic scale in 12-TET. Seven notes of 12-TET may be generated by one of two generators – 1 degree / 11 degrees,

or 5 degrees / 7 degrees. The smallest number of generators of 1 or 11 degrees required to generate the diatonic scale is 9, given that the smallest range that covers all seven notes of the scale is 10 degrees. We already know that the diatonic scale can be generated in 12-TET through stacking 12-TET's best perfect 5th of 7 degrees six times above (or below) a starting note. 12-TET's perfect 4th of 5 degrees may also be used, as it is the inversion of the perfect fifth and will therefore generate the same scales. The diatonic scale in 12-TET therefore has a SGC of 6. It follows that the SGC of any generated scale of  $n$  notes is  $n - 1$ . Carey (2002) suggests that both the minimum number of different generators for which it may be considered a generated scale (for which we were unable to build an algorithm) and the acoustic dissonance of the generators affects its scale candidacy.

### 5.3.5 Consonance

Consonance has received more definitions than there are researchers who write about it. We do not wish to give any definition of consonance, but to simply observe that the diatonic scale contains the highest number of triads and dyads generally considered to be consonant (e.g., perfect intervals, major and minor thirds and sixths, major and minor triads) out of any 7-note scale in 12-TET.

In tonal-harmonic music the tonic function belongs not only to a note but to a consonant triad (either major or minor). Major and minor triads are *tertian* in the diatonic scale, meaning that above the notes are separated by thirds in the scale. The previous chapter details an experiment run to test for perceived intrinsic stability of all possible triads of 22-TET, which we, here, take to be equivalent to consonance. Added to our analysis are measures of the maximum, median and minimum perceived stability for all tertian triads of each scale.

### 5.3.6 Tetrachordality

In terms of dyads, we assume that in 22-TET, as in 12-TET, the perfect fifth remains the strongest consonance (other than the octave). A mode of a scale is said to be *tetrachordal* if (in a single octave) it consists of two identical non-overlapping tetrachords that span an approximation of  $4/3$  (along with, necessarily, a step of an approximation of  $9/8$  as a remainder). Erlich (1998) defined a *tetrachordal* scale as a scale all of whose modes are tetrachordal. Such scales are now referred to more clearly as



*omnitetrachordal* (Erlich, 2017). We define tetrachordality as the number of modes of a scale of  $N$  notes that are tetrachordal, divided by the total number of modes,  $N$ . Tetrachordality therefore combines consonance (maximizing the number of perfect fifths and fourths) with redundancy (maximising the self-similarity of the scale at these intervals).

The diatonic scale is the only omnitetrachordal scale in 12-TET (giving it a tetrachordality of 1). In the diatonic scale, two large steps (of 2 degrees) and one small step (of 1 degree) make a perfect fourth (of 5 degrees) closely approximating  $4/3$ . The tetrachords that exist in the diatonic scale 2212221 are 221, 212 and 122. Placing two of these tetrachords in an octave leaves a 2-degree large step approximating  $9/8$  remaining. All modes of the diatonic scale are expressed below from brightest to darkest as two tetrachords and a remaining  $9/8$  step:

- Mode 3: 2 221 221
- Mode 2: 221 2 221
- Mode 1: 221 221 2
- Mode 0: 212 2 212
- Mode  $-1$ : 212 212 2
- Mode  $-2$ : 122 2 122
- Mode  $-3$ : 122 122 2

The modes are labelled by their *mode height*, as a measure of their brightness. The mode height is calculated as the deviation in degrees of the ET of the average pitch height of the mode from the middle of the octave. For a full definition and worked example, see Section 6.5.4 in Chapter 6.

The major mode, then, is Mode 2. It is the second brightest mode, after Mode 3, the Lydian mode – 2221221. The *mirror inversion* of Mode 3: 1221222, the Locrian mode, is Mode  $-3$ . Mode 0, Dorian mode – 2122212 – is *symmetric*: It is its own mirror inversion.

The WF diatonic scale of 22-TET – 4441441 – is also omnitetrachordal. This mode, the Lydian mode, has a mode height of 9. The modes for the WF diatonic scale in 22-TET are as such:

- Mode 9: 4 441 441
- Mode 6: 441 4 441

- Mode 3: 441 441 4
- Mode 0: 414 4 414
- Mode  $-3$ : 414 414 4
- Mode  $-6$ : 144 4 144
- Mode  $-9$ : 144 144 4

The pairwise well-formed (PWF) diatonic scale (the JI major scale) introduced in Section 5 above, with steps of LmsLmLs, is not omnitrachordal. This scale is also supported (approximated) by 22-TET, with  $(L, m, s) = (4, 3, 2)$  degrees. Inspecting its modes, from brightest to darkest again, with the tetrachordal modes (Modes 1,  $-2$  and  $-5$ ) expressed as two tetrachords and a conjunction (remainder), we have:

- Mode 6: 4342432
- Mode 3: 4324342
- Mode 1: 4 243 243
- Mode 0: 3424324
- Mode  $-2$ : 243 4 243
- Mode  $-3$ : 3243424
- Mode  $-5$ : 243 243 4

Unlike the WF diatonic scale (and all WF scales), this scale is not *mirror symmetric*, i.e., all of its modes cannot be expressed as the mirror inversion of another of its modes. The mirror inverse of this scale is 22-TET's approximation of the JI minor scale, with mode heights of  $-6$ ,  $-3$ ,  $-1$ ,  $0$ ,  $2$ ,  $3$  and  $5$ , with mode  $-6$  of this scale the mirror inverse of mode 6 of the 22-TET's approximation of the JI major scale, for example.

The tetrachordality of this mirror pair of scales is  $3/7$ .

## 5.4 Analysis

In order of mention, our features for analysis, according to their classification, are:

- Redundancy:
  1. mean variety
  2. maximum variety
  3. trivalence
  4. well-formedness
  5. pairwise well-formedness
  6. strong trivalence
  7. sameness quotient
  8. bichord entropy
  9. trichord entropy
  10. tetrachord entropy
  11. pentachord entropy
  12. hexachord entropy
- Coherence and evenness:
  13. coherence quotient
  14. Lumma stability
  15. Lumma impropriety
  16. evenness
- $R$ -ad entropy:
  17. dyad entropy
  18. triad entropy
- Generator complexity:
  19. scalar Graham complexity
- Consonance:
  20. min consonance
  21. max consonance
  22. median consonance
- Tetrachordality:
  23. tetrachordality

### 5.4.1 Reduction

We assume that, especially given the classification of these features into 6 groups, many may not be linearly independent of each other. Twenty-three is also a large number of features to consider in a cluster analysis and so we reduce the number of features. We do not use dimensional reduction, opting instead to select a subset of features that are least able to be expressed as linear combinations of the others in order that the set still comprises our original features rather than linear combinations of them. This means we can more directly test for the extent to which these features mediate the ability of a scale to support harmonic tonality. Such a test is undertaken in an exploratory analysis of the results of Experiment 4, detailed in Chapter 5. The features are first calculated for every 7-note scale in 22-TET. The *variance inflation factor*, or *VIF* (the factor by which the variance of a predictor is inflated compared to what you would expect if there was no multicollinearity; no correlation between predictors) is calculated for all the features, measuring the extent to which they may

be predicted by a linear combination of the other features. The VIF of a predictor can be calculated as

$$VIF = 1 - \frac{1}{1 - R^2} \tag{5.2}$$

using the  $R^2$  value of the linear regression of that predictor on all other predictors. The feature with the highest variance inflation factor is removed, and the process is iterated until the variance inflation factor for all remaining features is less than 2.

We found immediately that some of our features correlated 100% with each other: Hexachord entropy had only two values, depending on whether or not the scale was WF. It might be worth looking into  $n$ -chord entropy then, in future work, as a generalisation of well-formedness. Strong trivalence, we found, correlated 100% with trivalence. Where strong trivalence did not correlate 100% with pairwise well-formedness we have disproven Carey’s conjecture (introduced in Section 5.3.1): For example, 4334332 is an example of a strongly trivalent scale that is not pairwise well-formed. Though Carey proves by example that not all trivalent scales are strongly trivalent, we found that all trivalent 7-note scales in 22-TET are strongly trivalent. Removing hexachord entropy and strong trivalence first, our procedure leads us to the following features:

- Redundancy:
  1. maximum variety
  2. well-formedness
  3. pairwise well-formedness
  4. trichord entropy
  5. pentachord entropy
- Coherence and evenness:
  6. Lumma stability
  7. Lumma impropriety
- $R$ -ad entropy:
  8. triad entropy
- Generator complexity:
  9. scalar Graham complexity
- Consonance:
  10. min consonance
  11. max consonance
  12. median consonance
- Tetrachordality
  13. tetrachordality

The feature of evenness, we suspect, is captured, along with coherence, in Lumma stability and impropriety, given that they involve direct measures of relative interval size.

For use in Tables 5.2, 5.3 and 5.4, the above set of features are labelled and ordered as follows:

(9) SGC, (8) triad ent, (1) max var, (2) WF, (3) PWF, (4) 3chord ent, (5) 5chord ent, (6) Lumma stablty, (7) Lumma imprty, (10) max cos, (11) min cos, (12) med cos, (13) 4chrd-lty.

### 5.4.2 Cluster Analysis

Considering that our features are of different types of values – binary and continuous – we use Mahalanobis distance as our distance measure for our clustering. *K-medoids* clustering is used (via the *Partitioning Around Medoids* function in *R*) rather than *K-means* clustering given that exemplar scales from the original set are needed (*K-means* clustering would not give us actual scales as representatives of each cluster).

In order to test for the appropriateness of different numbers of clusters, we measure the average silhouette width for each clustering. The silhouette width for a single object is a measure of how similar it is to the cluster to which it is assigned, compared to the other clusters. It ranges from  $-1$  to  $1$ , where a high value indicates that the object is well classified in its cluster and a value below  $0$  indicates it is closer to another cluster, and may be misclassified (Rousseeuw, 1987).

The clustering algorithm leads us to a maximum at 9 clusters of our 7752 scales, with an average silhouette width of  $0.26$ . We observed however that the average silhouette width, and therefore the clustering may be substantially improved by leaving the vast majority of scales in a single cluster rather than splitting them into multiple clusters. Accordingly, from the initial clustering solutions for 3 to 40 clusters we combined clusters that appeared in plots of a principal components analysis (PCA) of the scales given the values of the 13 features to be less separated than other clusters such that the average silhouette width most improved. Further, misclassified scales (those with negative silhouette width) were moved into the cluster they are closest to when appropriate. Via these processes, we find a maximum average silhouette width of  $0.9877$  at 2 clusters, where one cluster is the scale 76 (4441441) and the other cluster is every other scale. We know from this that the scale 4441441 is the most distinct scale in terms of our features. Three clusters give the second best solution, consisting of 4441441, Scale 1(4333333) and the remainder, with average silhouette value  $0.9857$ . Following this, the other 5 well-formed scales split

from the remainder group as a cluster (for an average silhouette width of 0.9806), followed by scales 50 and 32 (for an average silhouette value of 0.9729) followed by Scale 11 (4342432) and its mirror inverse, Scale 13 (4342342) the pairwise well-formed (PWF) JI major scale (for an average silhouette width of 0.9606). The average silhouette width decreases incrementally for each larger number of new clusters until 12 clusters, in which the decrement from 11 clusters is substantially larger (0.9018 to 0.7646).

Figures 5.1 and 5.2 show rotations of a plot of the clustering in the first three principal components from a 3D PCA (which account for 22%, 18% and 10% of the variance respectively), with the scales labelled 1-7752, ordered from most to least even. For interpretability, the representation of the 13 features in the principal components are plotted as vectors with labels at 15 standard deviations from the origin, though PWF (pairwise well-formedness) is mostly hidden (you can kind of see it in the **maroon** cluster, which comprises PWF scales). Eleven clusters seems quite appropriate looking at the clustering, and 11 scales is already pushing towards, or possibly through the limit on how many scales we can test in an experiment. Accordingly, we take 11 clusters to be a stopping point. For supplementary material, including an interactive 3D PCA plot of the clustering, data for all 7752 scales, and sound files for the exemplar scales, follow this link: <https://en.xen.wiki/w/User:Gareth.hearne/Analysis22-7>

### 5.4.3 Exemplar Scales

Table 5.1 displays the exemplar scales in hexadecimal, along with their scale ID, so they can be located in the cluster diagram. They are ordered such that the first  $n$  scales are the exemplars for the best  $n$ -cluster solution, and the size of each new cluster, and the average silhouette value for each associated successive clustering is also shown. As the first clustering solution is into 2 clusters – and – clusters, cluster size for the cluster is written as 'NA'. It does not make sense to have a single cluster clustering solution.

Tables 5.2 and 5.3 display the values of the 13 features and their Z-scores for these scales.

The scale 8113621 represents the vast majority of scales (**magenta**), for which all features are valued within 1 standard deviation of the mean. The scale 4441441 (**pink**), the WF scale generated by the approximation of  $3/2$ , is the most exceptional

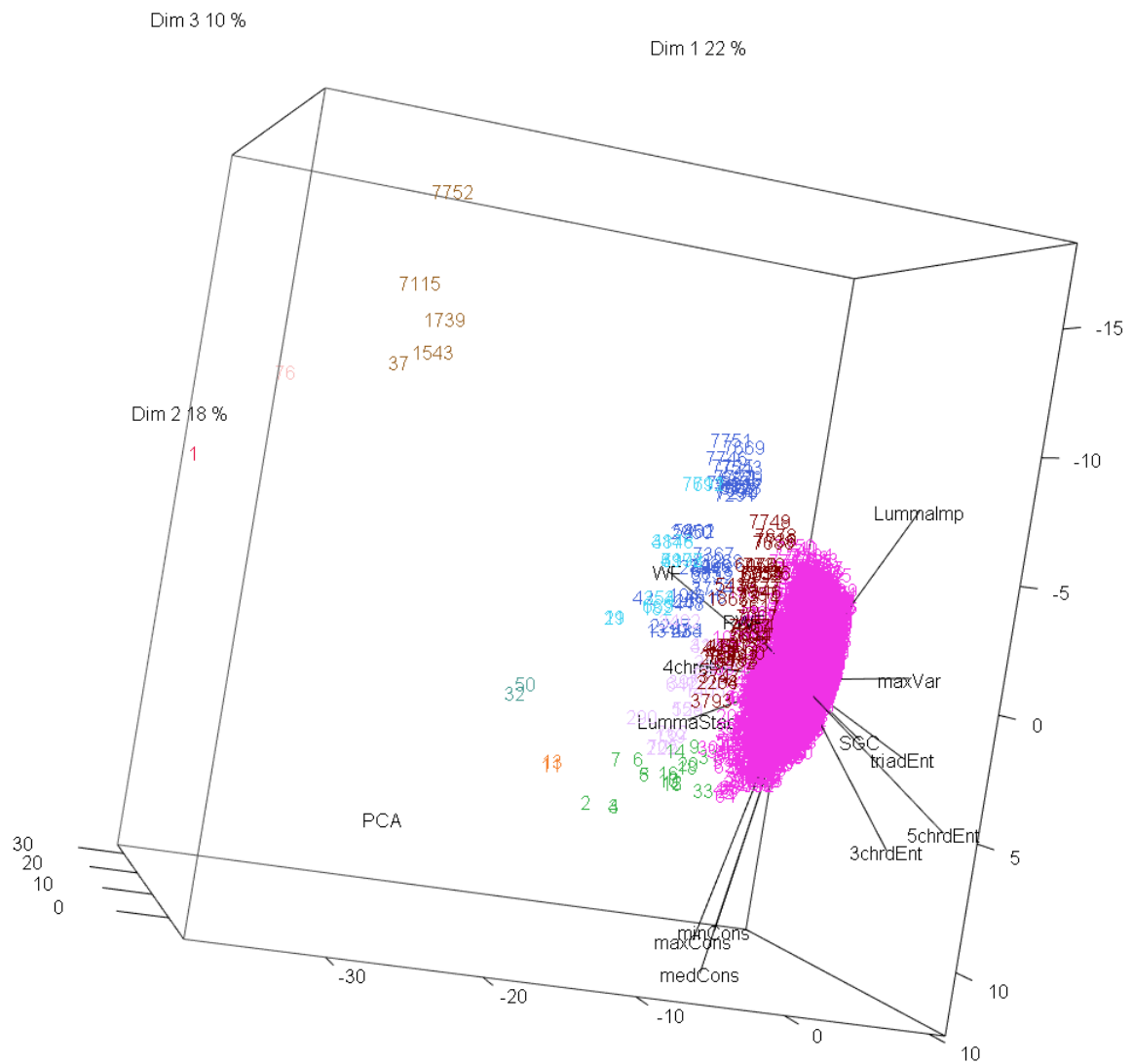


FIGURE 5.1: 3D clustering view 1

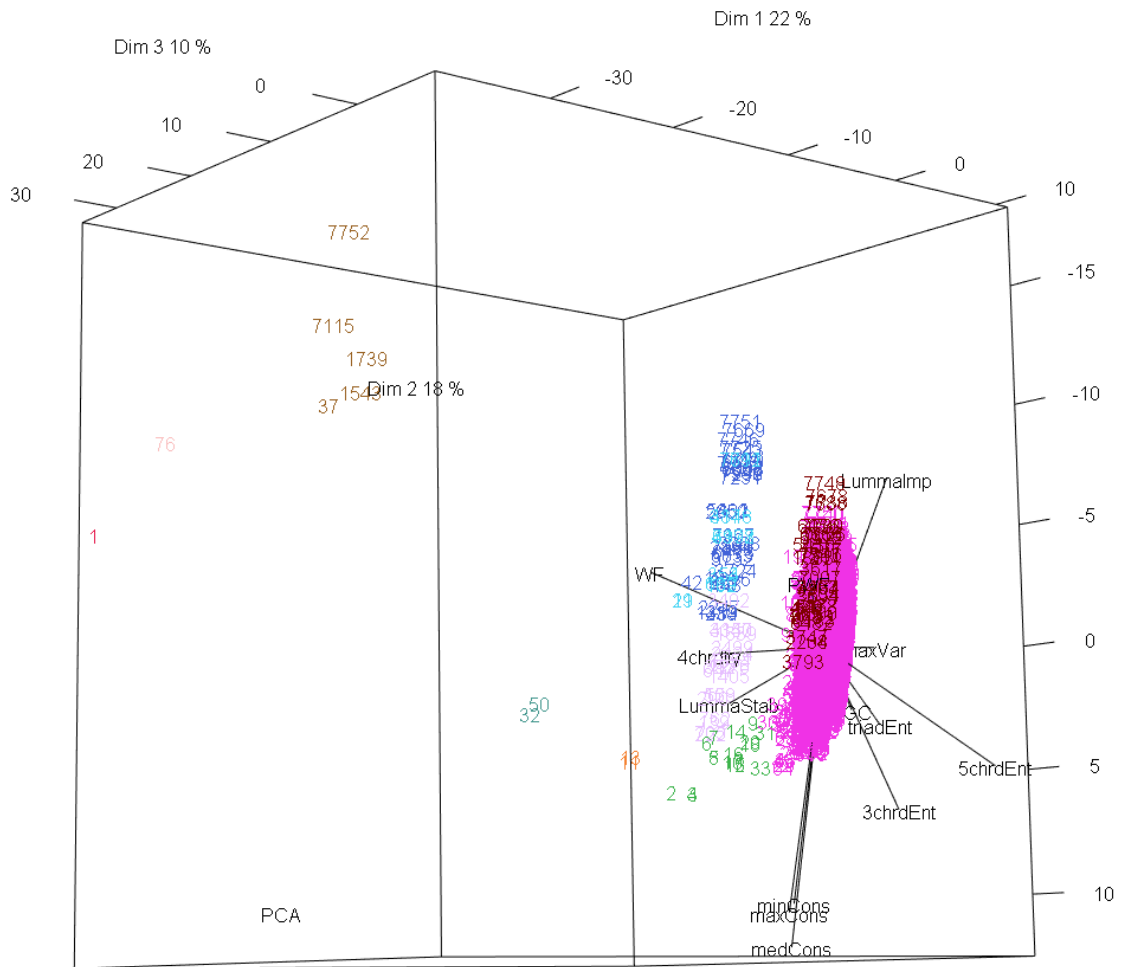


FIGURE 5.2: 3D clustering view 2



TABLE 5.1: Exemplar scales associated with each successive cluster added.

Number of clusters	Added cluster	Exemplar Scale ID	Exemplar Scale	Cluster size	Average silhouette width
NA	magenta	4866	8113621	NA	NA
2	pink	76	4441441	1	0.9877
3	red	1	4333333	1	0.9857
4	brown	1739	6226222	5	0.9806
5	teal	50	4432432	2	0.9729
6	orange	13	4342432	2	0.9606
7	green	17	4343332	18	0.9374
8	lavender	1405	6142612	18	0.9340
9	cyan	4397	7414141	14	0.9243
10	blue	7367	B122222	34	0.9068
11	maroon	4954	8121811	40	0.9018

TABLE 5.2: Values of features for exemplar scales.

Scale	SGC	triad ent	max var	WF	PWF	3chrd ent	5chrd ent	Lumma stablty	Lumma imprty	max cons	min cons	med cons	4chrd -lty
8113621	12	7.26	5	0	0	2.81	2.81	0	1	0.54	1.90	1.03	0
4441441	6	6.29	2	1	0	1.56	2.24	0	5/22	0.74	1.96	1.25	1
4333333	6	6.29	2	1	0	1.15	2.13	8/11	0	0.95	3.23	1.83	3/7
6226222	6	6.23	2	1	0	1.56	2.24	0	1	0.80	1.07	0.81	0
4432432	11	7.12	4	0	0	1.95	2.52	2/11	2/11	1.04	3.23	1.83	5/7
4342432	11	7.04	3	0	1	2.24	2.81	9/22	1/22	1.04	3.23	1.83	3/7
4343332	11	7.31	4	0	0	2.24	2.81	3/11	0	0.95	3.23	1.43	0
6142612	11	7.37	5	0	0	2.24	2.81	0	21/22	0.80	3.23	1.25	3/7
7414141	11	7.04	3	0	1	1.84	2.52	0	1	0.74	1.96	1.17	0
B122222	11	7.04	4	0	0	1.66	2.52	0	1	0.39	1.79	0.86	0
8121811	12	6.99	3	0	1	2.24	2.81	0	1	0.54	1.90	1.03	0

TABLE 5.3: Z-scores of features for exemplar scales.

Scale	SGC	triad ent	max var	WF	PWF	3chrd ent	5chrd ent	Lumma stablty	Lumma imprty	max cons	min cons	med cons	4chrd -lty
8113621	-0.13	-0.79	-0.71	-0.03	-0.09	0.68	0.08	-0.07	0.40	0.17	-0.13	-0.11	-0.05
4441441	-4.00	-6.09	-4.15	33.26	-0.09	-5.44	-19.31	-0.07	-3.69	0.72	-0.05	0.72	36.31
4333333	-4.00	-6.09	-4.15	33.26	-0.09	-7.43	-22.97	30.40	-4.90	1.29	1.71	2.92	15.53
6226222	-4.00	-6.40	-4.15	33.26	-0.09	-5.44	-19.31	-0.07	0.40	0.88	-1.29	-0.94	-0.05
4432432	-0.77	-1.54	-1.86	-0.03	-0.09	-3.51	-9.62	7.55	-3.93	1.54	1.71	2.92	25.92
4342432	-0.77	-1.97	-3.01	-0.03	11.42	-2.11	0.08	17.07	-4.66	1.54	1.71	2.92	15.53
4343332	-0.77	-0.48	-1.86	-0.03	-0.09	-2.11	0.08	15.16	-4.90	1.29	1.71	1.40	-0.05
6142612	-0.77	-0.17	-0.71	-0.03	-0.09	-2.11	0.08	-0.07	0.16	0.88	1.71	0.72	15.53
7414141	-0.77	-1.97	-3.01	-0.03	11.42	-4.04	-9.62	-0.07	0.40	0.72	-0.05	0.42	-0.05
B122222	-0.77	-1.97	-1.86	-0.03	-0.09	-4.91	-9.62	-0.07	0.40	-0.24	-0.29	-0.76	-0.05
8121811	-0.13	-2.28	-3.01	-0.03	11.42	-2.11	0.08	-0.07	0.40	0.17	-0.13	-0.11	-0.05

(and probably the most similar to 12-TET's diatonic scale), and the scale 4333333 (red), the maximally even scale, the second most exceptional. The scale 6226222 represents the other 5 WF scales (brown). 4432432 represents itself and its mirror inverse 4423423 (teal), the two scales with tetrachordality of 5/7. 4342432 represents itself and its mirror inverse 4342342 (orange), the PWF scales with tetrachordality value 3/7. 4343332 represents the remaining scales that are relatively consonant, with Lumma stability above 0 and low Lumma impropriety (green). 6142612 represents the other scales with tetrachordality 3/7 (lavender). The remaining PWF scales are split between the clusters represented by 7414141 (cyan), and 8121811 (maroon), the first being those with pentachord entropy 2.52, and the second, 2.81, which is very close to the mean for all scales. The final exemplar scale B122222 represents the scales with pentachord entropy 2.52 that are not PWF (blue) (Figs. 5.1 and 5.2).

The clustering seems to be dominated by WF, PWF and tetrachordality, the variables for which the few possible values other than 0 are very rare. This is probably because extreme values of these features can cause scales to “stand out” more overall than extreme values of other features.

We note that scales 1, 13, 50 and 76 can be thought of as 22-TET's approximations of 4 different JI representations of the diatonic scale. We'll begin with Scale 13, 4342432, which, in its mode 4324342, is 22-TET's approximation of the JI major scale,  $9/8$   $5/4$   $4/3$   $3/2$   $5/3$   $15/8$   $2/1$ , which is PWF. Scale 50 is very similar. In its mode 4324432 it is 22-TET's approximation of an alternative JI major scale,  $9/8$   $5/4$   $4/3$   $3/2$   $27/16$   $15/8$   $2/1$ . This scale is not PWF, but it has tetrachordality 5/7, rather than 3/7. If we take its steps to be of 22 (unequal) *śruti* of early Indian music, rather than of degrees of 22-TET, these two scales are (modes of) the two basic scales of early Indian music, *Ma grāma* and *Sa grāma*, respectively (Clough et al., 1993; Daniélou, 1995; Erlich, 1998). A third scale, '*Ga grāma*', though less frequently discussed, also existed. Though the tuning is quoted differently across sources, Daniélou (1995) suggests that it is 3334333, which in 22-TET is Scale 1, 22-TET's approximation of the PWF JI Dorian scale  $10/9$   $6/5$   $4/3$   $3/2$   $5/3$   $9/5$   $2/1$ .

Finally, Scale 76 in its mode 4414441 we already know is 22-TET's approximation of the Pythagorean diatonic scale  $9/8$   $81/64$   $4/3$   $3/2$   $27/16$   $243/128$   $2/1$ . It can also be thought of as approximating the scale  $9/8$   $9/7$   $4/3$   $3/2$   $27/16$   $27/14$   $2/1$ , in a similar way to how in 12-TET the scale 2212221 approximates both Pythagorean and JI major scales.

The last two scales (4414441 and 4333333), the most distinct in 22-TET, are probably the most popular among musicians who use 22-TET, referred to as ‘Superpyth[7]’ and ‘Porcupine[7]’ respectively. This analysis suggests we should not be surprised by this.

## 5.5 Conclusion

After a review of scale features that may relate to the cognition of tonality, a distributional analysis of all 7-notes scale of 22-TET in terms of their values for these features led to the selection of 11 exemplar scales representative of the entire set, for use in a pair of experiments in the next chapter. As it happened, however, Experiments 4 and 5, detailed in the following chapter, had commenced using a selection of 3936 scales obtained from a previous analysis that erroneously considered mirror pairs of scales to be a single scale before the error was discovered, and were left to run their course, considering time constraints. In terms of our features, a mirror pair of scales differ only in their consonance values. The 8 scales used for the analysis contained 6 of the 11 scales we arrived at in this chapter – 8113621, 4441441, 4333333, 6226222, 4342432, 6142612 and 8113621 – along with two additional scales – 4343242 and B116111. 4343242 (Scale 15) replaces 4343332 (Scale 17), representing the scales that are relatively consonant that are not WF or have tetrachordality values above 0, with Lumma stability above 0 and low Lumma impropriety (**green**). B116111 (Scale 7346) represents the PWF scales, apart from the mirror pair represented by 4342432, replacing both 7414141 (Scale 4397) and 8121121 (Scale 4954), combining the clusters represented by these two scales together. Missing from the set used in the experiments in Chapter 6 are 4432432 and its mirror pair (**teal**), which are represented in the scales used in Chapter 6 by 6142612 with the cluster of non WF scales with tetrachordality values above 0; and the scales with pentachord entropy 2.52 that are not PWF (**blue**), which had been represented by B122222, absorbed in the set of scales used in Chapter 6 into the cluster of the majority of scales. The scale 8113621 represents the vast majority of scales (**magenta**). The values of the features of the scales 4343242 and B116111 are shown in Table 5.4.

Though Experiments 4 and 5 do not use the optimal set of scales, the set used did not differ too much from the optimal set derived in this chapter, and we do not expect the results to differ a meaningful amount.

TABLE 5.4: Values of features for exemplar scales.

Scale	SGC	triad ent	max var	WF	PWF	3chrd ent	5chrd ent	Lumma stably	Lumma imprty	max cons	min cons	med cons	4chrd -lty
4343242	11	7.4857	4	0	0	2.2359	2.8074	4/11	0	0.95	3.23	1.44	0
B116111	11	7.0425	3	0	1	2.1281	2.8074	0	2 1/22	0.37	1.46	0.71	0

## Chapter 6

# Experiments 4 & 5: Stability of probe tones and triads in microtonal scales

### 6.1 Introduction

A pair of experiments (Experiments 4 and 5) were conducted to test the perceived stability of tones and tertian triads after the context of an isochronous, randomly ordered sounding of the pitch classes of 8 microtonal 7-note scales from 22-TET (the 8 scales come out of an earlier cluster analysis for which 8 was a clear stopping point for the number of clusters). Tonal hierarchies were found in all scales bar one for the probe tone experiment, and bar two for the probe triad experiment. For the probe tone experiment a single RPC was rated clearly highest for scales that approximate the 12-TET diatonic scale; for the probe triad experiment a single chord was rated clearly highest for the scales which contained at least one triad approximating the justly tuned major triad. We suggest that these RPCs and triads may function well as tonics in tonal-harmonic compositions. Bayesian ordinal mixed effects models of the results from both experiments were significantly weakened by the removal of SPCS, suggesting that SPCS is indeed able to predict ratings of tones and triads in novel, microtonal scales. Exploratory analyses suggest that SPCS predicts more detail in ratings than simply whether or not the probe was heard in context, and accordingly, models including SPCS perform better than models that instead include ScaleTone. It is speculated that the model may be improved by the addition of a number of effects observed in a descriptive analysis of the data. Further research could be conducted to test for the effectiveness of such a model in predicting stability ratings of probe tones in novel microtonal scales.

## 6.2 Hypotheses

H1: The perceived stability of the probe may be modelled by the SPCS between the pitches of the context and the probe and, for Experiment 5, the intrinsic perceived stability of the probe triad type.

H2: Significant interaction effects exist between musical sophistication and other predictors in such a model.

## 6.3 Method

### 6.3.1 Participants

#### Experiment 4: Tones

Twenty-four musicians (participants who reported having received 5 or more years of music experience) and 36 non-musicians were recruited for the experiment. The data from 1 non-musician and 3 musician participants were lost. The following refers to the 56 remaining participants. Non-musician participants were university students (mostly first-year) recruited through Western Sydney University School of Psychology and Social Science's SONA system and received credit points towards their degrees for their participation. Musician participants were recruited via personal connection and received a \$30 reimbursement for their time and travel to the university campus. All participants demonstrated normal hearing capabilities. Fifteen (musician) participants reported having received 10 or more years of musical training and one reported having absolute pitch. Participants had a mean age of 25.1 years, with a SD of 8.9 years. Of the 21 musicians, 7 were female, and of the 35 non-musicians, 28 were female. This research was approved by the Western Sydney University Human Research Ethics Committee under the number H11908.

#### Experiment 5: Triads

Twenty-four musicians (participants who reported having received 5 or more years of music experience) and 36 non-musicians were recruited for the experiment. Non-musician participants were university students (mostly first-year) recruited through

Western Sydney University School of Psychology and Social Science's SONA system and received credit points towards their degrees for their participation. Musician participants were recruited via personal connection and received a \$30 reimbursement for their time and travel to the university campus. All participants demonstrated normal hearing capabilities. 11 (musician) participants reported having received 10 or more years of musical training and four reported having absolute pitch. Participants had a mean age of 24.1 years, with a SD of 7.2 years. Of the 24 musicians, 17 were female, and of the 36 non-musicians, 26 were female. This research was approved by the Western Sydney University Human Research Ethics Committee under the number H11908.

### 6.3.2 Stimuli

The stimuli differ from the stimuli of Experiments 1 and 2 only by the context scales and the probes.

The context scales for this pair of experiments were the 8 scales chosen via a cluster analysis considering all possible heptatonic scales in a tonal scale feature space, each scale being the medoid of a cluster in the solution. The reader will recall that Chapter 5 led to the selection of 11 such scales, rather than to 8. The 8 scales used in this experiment are the results of a previous clustering solution that was updated to what is shown in Chapter 5 after data collection had already begun. The previous cluster analysis was of a smaller set of scales in which, for example, only one scale from each mirror pair, e.g., the harmonic major and harmonic minor scales are included, these scales being equivalent in terms of all features other than consonance.

Described in their brightest modes, as step lists in degrees of 22-TET, the scales are

1. 4 3 3 3 3 3 3
2. 4 3 2 4 3 4 2
3. 4 3 4 3 2 4 2
4. 4 4 4 1 4 4 1
5. 6 2 2 6 2 2 2
6. 6 1 2 6 1 2 4
7. 8 1 1 3 6 2 1

## 8. B 1 1 6 1 1 1

Hexadecimal notation is used for scale steps of more than 9 degrees, where ‘B’ represents 11 degrees.

For the probe tone experiment (Experiment 4) each RPC of the scale was probed, along with a random selection of 7 of the remaining 15 RPCs (which are non-scale-tones).

For the probe triad experiment (Experiment 5) each tertian triad of the scale was probed, along with a random selection of probes rooted on 7 of the remaining 15 RPCs. These randomly selected triads were of the same type (i.e., transpositional and inversional and equivalence class, like major, minor, diminished and augmented for 12-TET) as the tertian triads in the scale, selected randomly with proportions equivalent to that of the triad types within the tertian triads of the scale. Each probe was heard twice by participants.

### 6.3.3 Procedure

For both experiments, as in Experiment 3 (detailed in Section 4.4.2), for each trial, participants were asked to rate the ‘stability of the final musical sound given the sounding of the context melody’, on a 7-point Likert scale after being informed that ‘a musical sound is considered to be stable if it does not need to move (resolve) to another music sound’.

For both experiments probes are randomly ordered with the sections for each scale, and the scales are randomly ordered. This results in  $(7 + 7) \times 2 \times 8 = 224$  trials for each experiment. The experimental trials were preceded by 8 practice trials – one in each scale – ordered randomly, independent of the order of scales in the experimental blocks. Participants rate stability as in Experiment 3. Half way through the experiment, participants completed a survey including the Goldsmith MSI Questionnaire in order to obtain an index for musical sophistication to be used as a variable in analysis. Additional demographic questions followed the Goldsmith MSI Questionnaire to facilitate future analysis of possible effects of enculturation, though these are not analysed here.



## 6.4 Analysis

As in Experiments 1 and 2, analysis includes a test for tonal hierarchies in each scale, a descriptive model, and both confirmatory and exploratory Bayesian ordinal mixed effect models. Significance is determined as in Experiments 1, 2 and 3.

To test for the emergence of tonal(-harmonic) hierarchies in our microtonal scales we compare two simple Bayesian ordinal (cumulative logit) regression models, the null and the alternative. In the alternative model, ratings are predicted only by the RPC of the probe tone (for Experiment 4) or by an interaction of the RPC of the probe triad's root and its observed stability, as collected from Experiment 3 (for Experiment 5). In both cases the associated null model uses only the intercept to predict ratings.

For both experiments, linear models were run before the ordinal mixed effects models that were used for confirmation of the hypotheses. For the probe tones experiment, SPCS was used to predict the observed ratings for each scale-probe combination, averaged over all other variables; for the probe triads, predictors comprised SPCS and the *intrinsic stability* of the probe triad, as observed in Experiment 3, averaged over all other variables.

Confirmatory analyses for each experiment involve Bayesian ordinal (cumulative logit) mixed effects models of all effects included in Experiment 3, but for Triad in Experiment 4, and in Experiment 5 with Triad replaced with *ChordStab*, a measure of the intrinsic stability of the probe triad as obtained from experiment 3. Hypothesis 1 is confirmed if a model comparison reveals the alternative model to be significantly stronger than the null, which is identical but for the absence of SPCS, for Experiment 4, and of both SPCS and *ChordStab* for Experiment 5. Hypothesis 2 is confirmed by the presence of significant interactions with musical sophistication.

Exploratory analyses follow, with one comparing the alternative models (Model Tone and Model Triad respectively) to models identical, but with *ScaleTone* replacing SPCS, another adding to Model Tone an effect of Scale (a categorical effect what which scale was heard in the context stimulus), another simplifying Model Tone and adding the 13 scale features of Chapter 4 as effects, and a final analysis adding to Model Tone, four additional effects suggested by observation of plots of the linear models.

As in previous Chapters, Bayesian mixed effects models are run in *brms* in *R*, using

$\text{student\_t}(3, 0, 2.5)$

(a  $t$ -distribution with 3 degrees of freedom, with mean of 0, scaled by 2.5) as a weakly informative prior. As for the analysis of Experiments 1 and 2, All continuous independent variables were *standardized* – centered at 0 and scaled to have a standard deviation of 1. Model comparisons are made via the leave-one-out cross validation information criterion (LOOIC), and are considered to differ significantly when the LOOIC differ by more than twice the SE associated with the comparison.

## 6.5 Results

### 6.5.1 Test for tonal hierarchies

For Experiment 4 (probe tones), hierarchies were uncovered for all scales bar Scale 5 – 6226222. The results of the LOOIC comparisons are shown in Table 6.1. It is worth noting that Scale 5 is the only scale tested that does not include the ETs approximation to a perfect fifth, which is the most consonant interval within an octave.

TABLE 6.1: LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 4

Scale	Null – Alternative LOOIC	SE	Signif
4333333	105.4	33.2	Yes
4324342	294.4	48.6	Yes
4343242	118.8	33.6	Yes
4441441	268.6	43.8	Yes
6226222	45.4	27.2	No
6126124	92.0	31.4	Yes
8113621	116.6	33.2	Yes
B116111	148.0	36.4	Yes

For Experiment 5 (probe triads), hierarchies were uncovered for all scales bar Scale 5 – 6226222, and Scale 8 – B116111. The results of the LOOIC comparisons are shown in Table 6.2. We note that these are the only two scales tested that do not include an approximation of a major triad (i.e., an interval within 50c of a 12-TET major third of 400c above the same pitch class on which there also sits the tuning’s best approximation of a perfect 5th).

TABLE 6.2: LOOIC comparisons of simple Bayesian ordinal mixed effects models for test of tonal hierarchy for the scales of Experiment 5

Scale	Null – Alternative LOOIC	SE	Signif
4333333	275.6	51.0	Yes
4324342	259.2	46.2	Yes
4343242	281.8	53.4	Yes
4441441	107.2	41.2	Yes
6226222	50.8	33.6	No
6126124	418.8	59.6	Yes
8113621	115.0	26.0	Yes
B116111	13.0	31.0	No

### 6.5.2 Descriptive Model

Figures 6.1 – 6.24 plot, for each scale, first the SPCS’s predicted stability ratings of each scale-tone probe RPC for each context scale against observed ratings; then predicted against observed ratings for all RPCs; and finally ratings predicted by SPCS and intrinsic stability of tertian probe triads against observed ratings for the probe triad experiment. For the plots of probe tone ratings, scale-tones are numbered as RPCs; for plots of probe triad ratings, triads types are labelled using the convention established in Section 4.6.1, with their roots labelled by their RPC (Scales 3 and 5–8), or, when the scale is similar enough to the diatonic scale, by the Roman numeral for their scale-tone. with the RPCs of the other pitches of the triad shown in brackets. Error bars are from 95% confidence intervals obtained from 1000 bootstrapped samples<sup>1</sup> In an attempt to simplify interpretability, the RPC of each scale that received the highest rating in the probe tone experiment is given the RPC value 0. Scales are then shown in the mode for which RPC 0 is the lowest pitch class, as in Chapters 2 and 3. For the diatonic-like scales – 4324342 and 4414441 – this corresponds to the major mode.

<sup>1</sup>Chromatic triads (triads including non-scale-tones) were included in the model but are not shown in the figures.

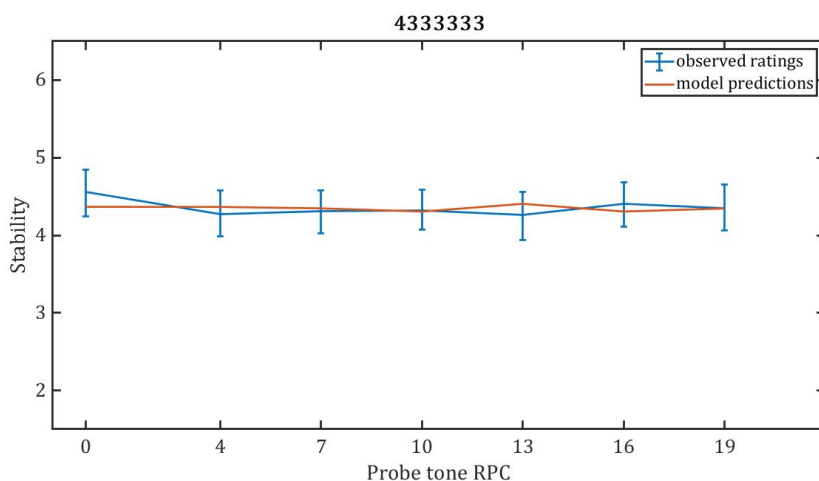


FIGURE 6.1: Average ratings for probe tones after a context of Scale 1, compared to SPCS predictions – scale-tones only, numbered as RPCs.

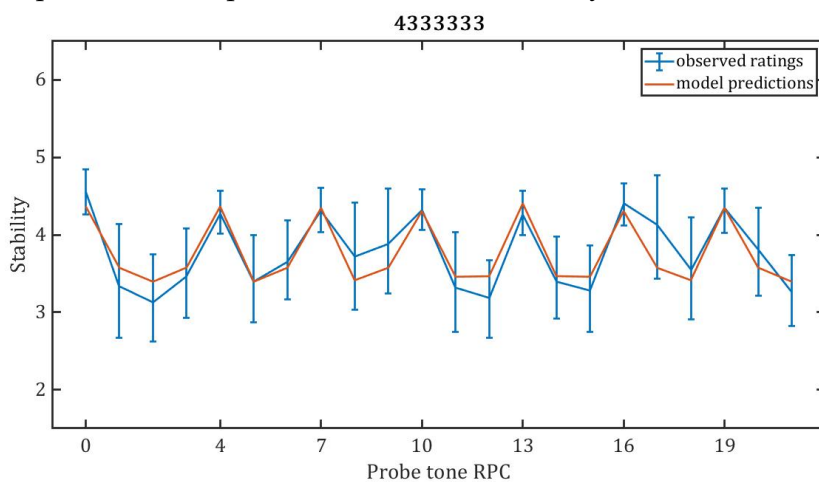


FIGURE 6.2: Average ratings for probe tones after a context of Scale 1, compared to SPCS predictions – all RPCs, with scale-tones numbered.

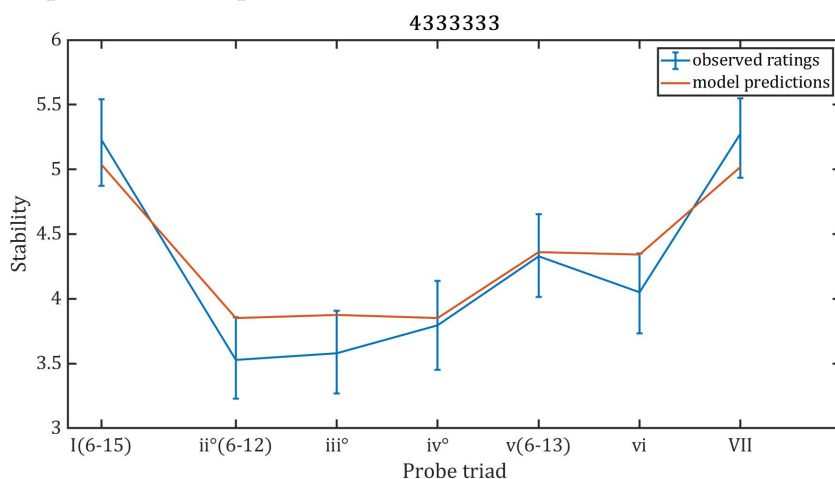


FIGURE 6.3: Average ratings for probe triads after a context of Scale 1, compared to SPCS predictions. Triads roots are labelled by Roman numeral.

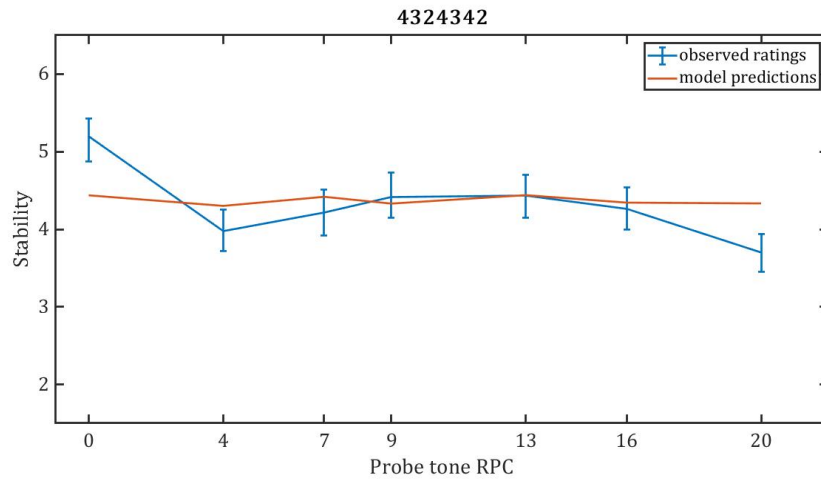


FIGURE 6.4: Average ratings for probe tones after a context of Scale 2, compared to SPCS predictions – scale-tones only.

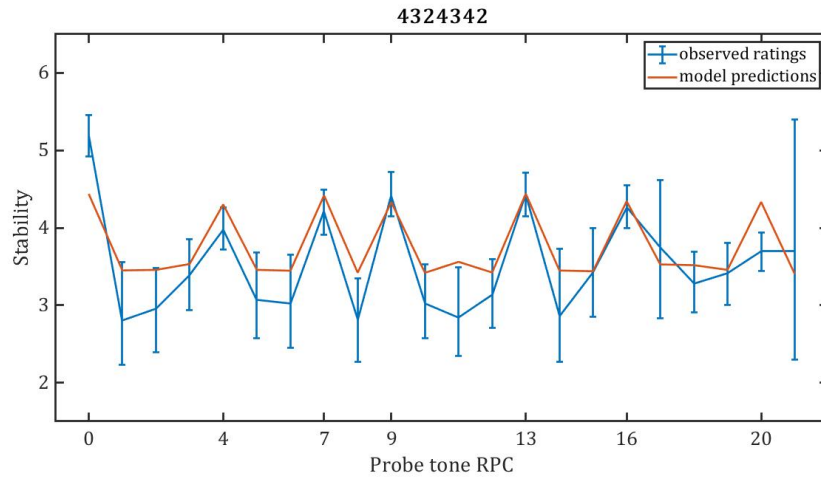


FIGURE 6.5: Average ratings for probe tones after a context of Scale 2, compared to SPCS predictions – all RPCs.

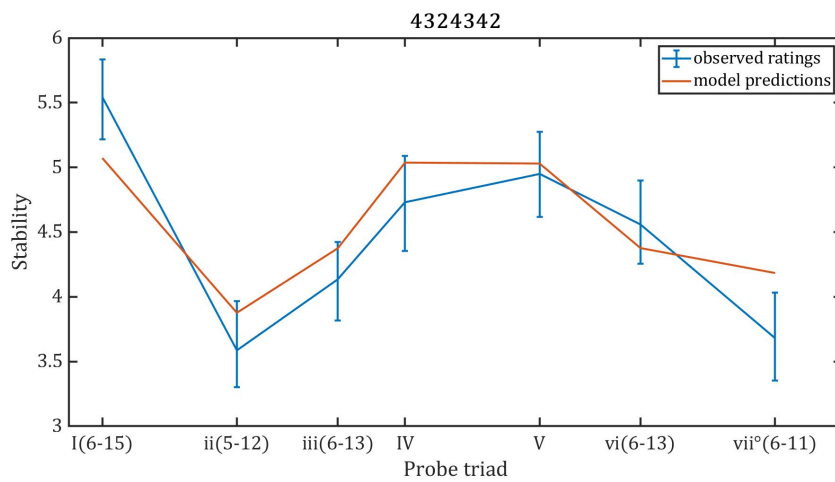


FIGURE 6.6: Average ratings for probe triads after a context of Scale 2, compared to SPCS predictions. Triads roots are labelled by Roman numeral.

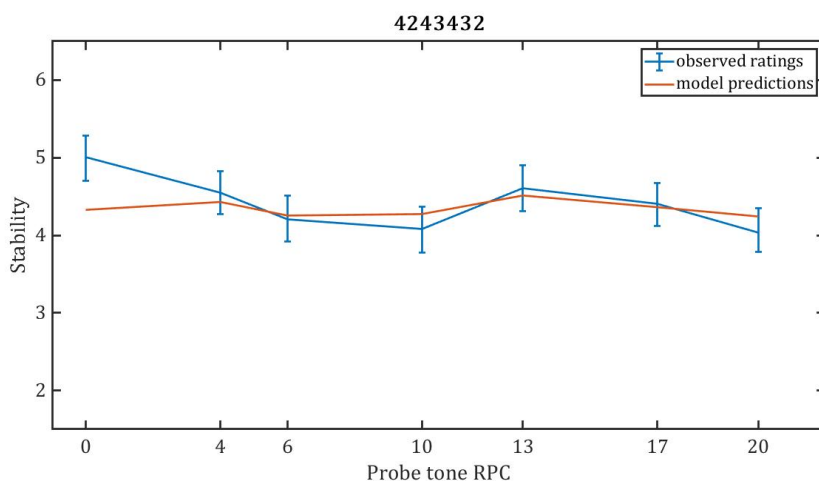


FIGURE 6.7: Average ratings for probe tones after a context of Scale 3, compared to SPCS predictions – scale-tones only.

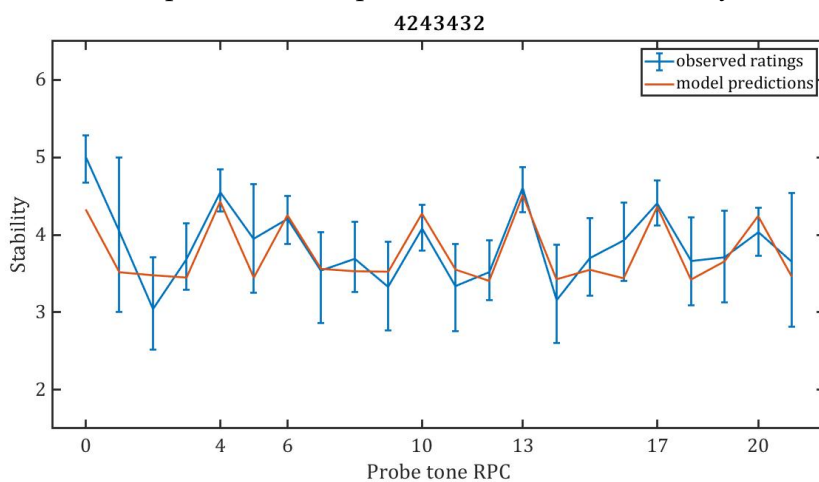


FIGURE 6.8: Average ratings for probe tones after a context of Scale 3, compared to SPCS predictions – all RPCs.

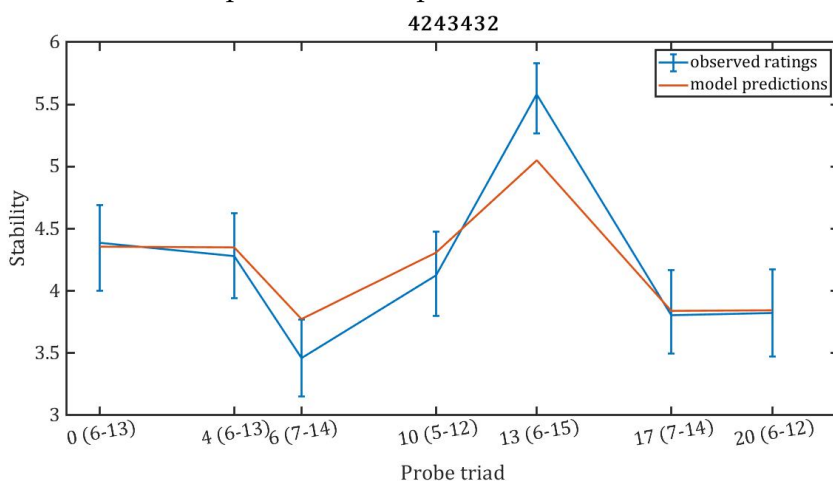


FIGURE 6.9: Average ratings for probe triads after a context of Scale 3, compared to SPCS predictions. Triads roots are labelled by RPC.

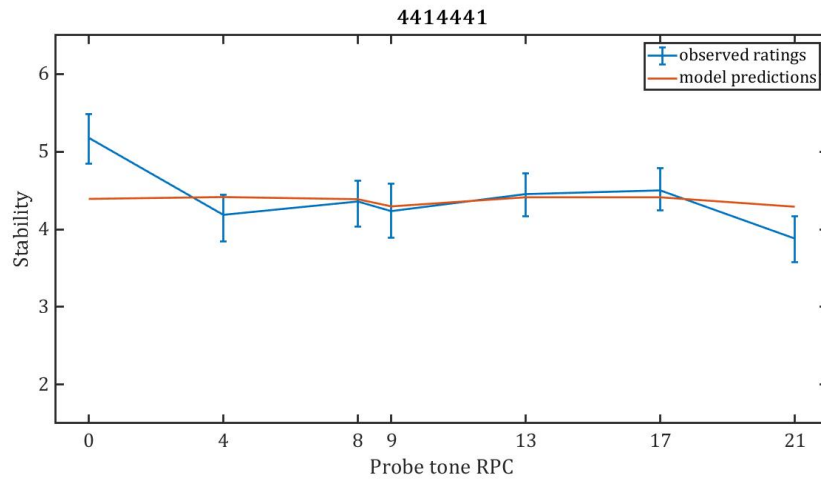


FIGURE 6.10: Average ratings for probe tones after a context of Scale 4, compared to SPCS predictions – scale-tones only.

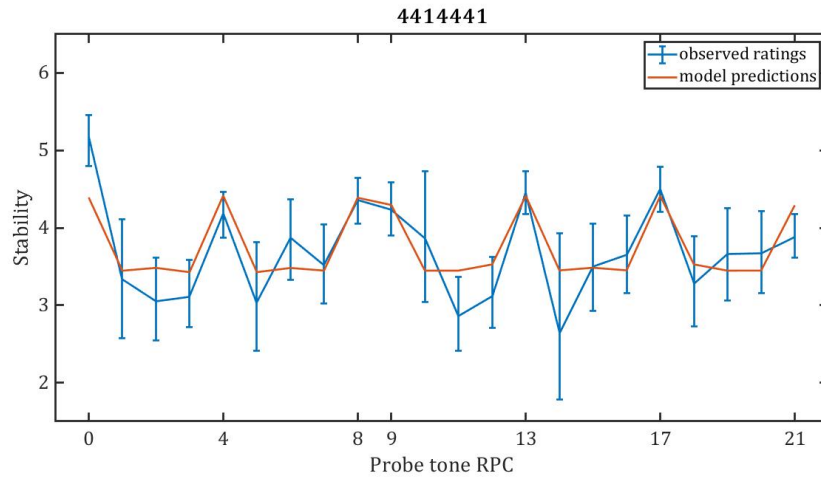


FIGURE 6.11: Average ratings for probe tones after a context of Scale 4, compared to SPCS predictions – all RPCs.

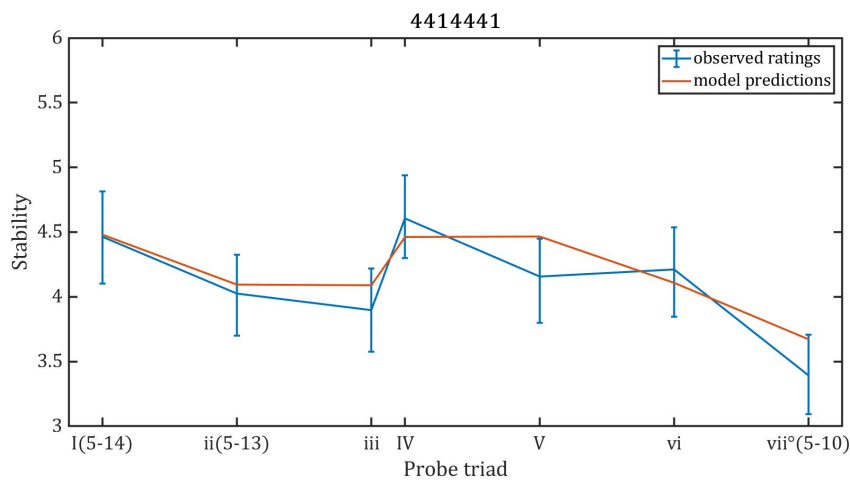


FIGURE 6.12: Average ratings for probe triads after a context of Scale 4, compared to SPCS predictions. Triads roots are labelled by Roman numeral.

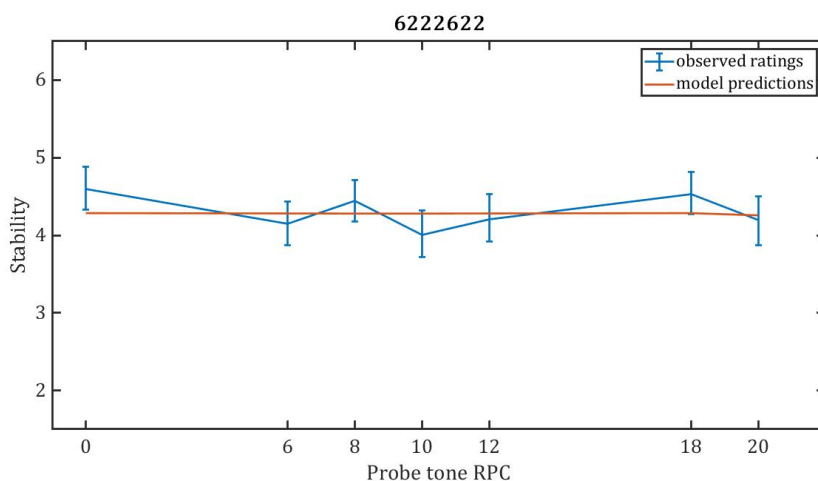


FIGURE 6.13: Average ratings for probe tones after a context of Scale 5, compared to SPCS predictions – scale-tones only.

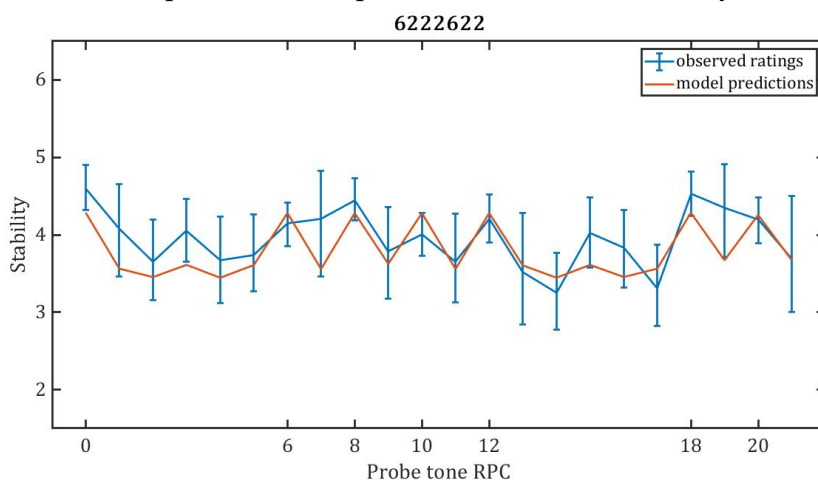


FIGURE 6.14: Average ratings for probe tones after a context of Scale 5, compared to SPCS predictions – all RPCs.

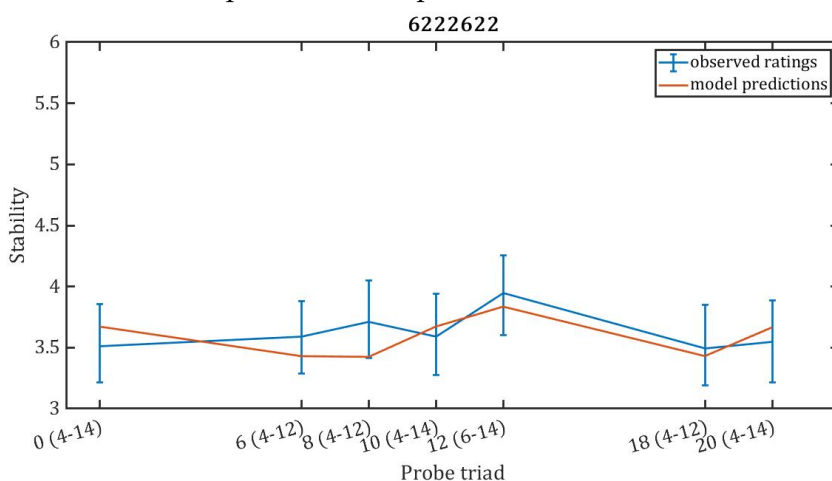


FIGURE 6.15: Average ratings for probe triads after a context of Scale 5, compared to SPCS predictions. Triad roots are labelled by RPC.



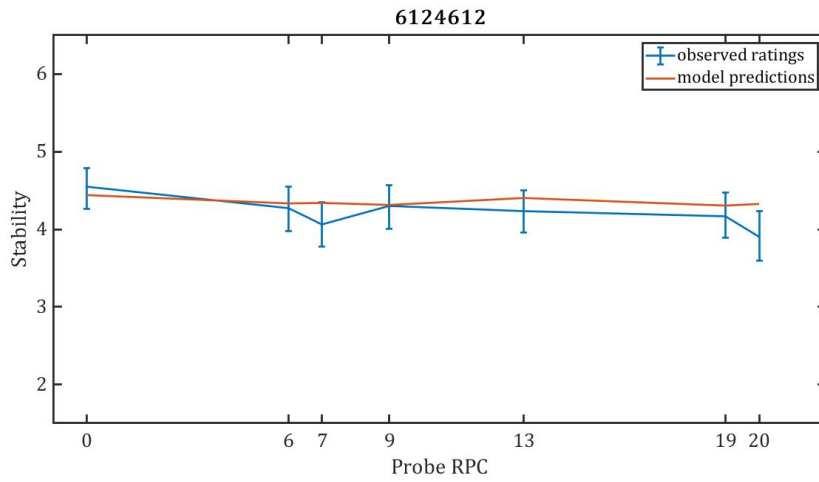


FIGURE 6.16: Average ratings for probe tones after a context of Scale 6, compared to SPCS predictions – scale-tones only.

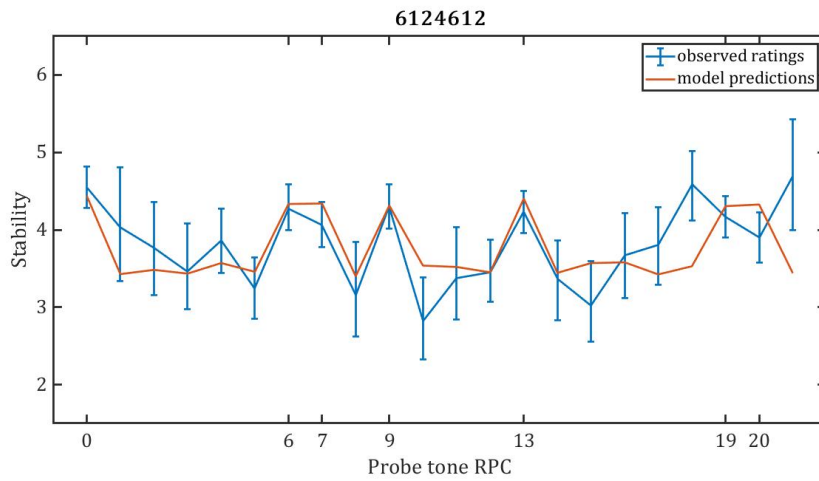


FIGURE 6.17: Average ratings for probe tones after a context of Scale 6, compared to SPCS predictions – all RPCs.

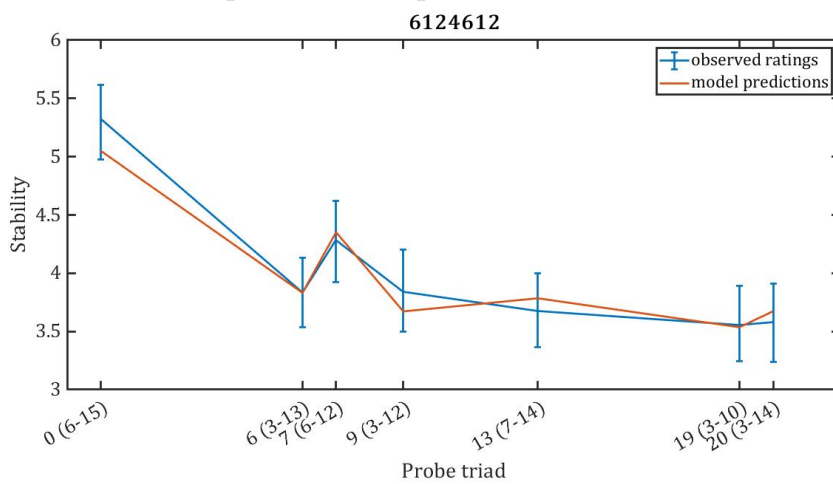


FIGURE 6.18: Average ratings for probe triads after a context of Scale 6, compared to SPCS predictions. Triad roots are labelled by RPC.

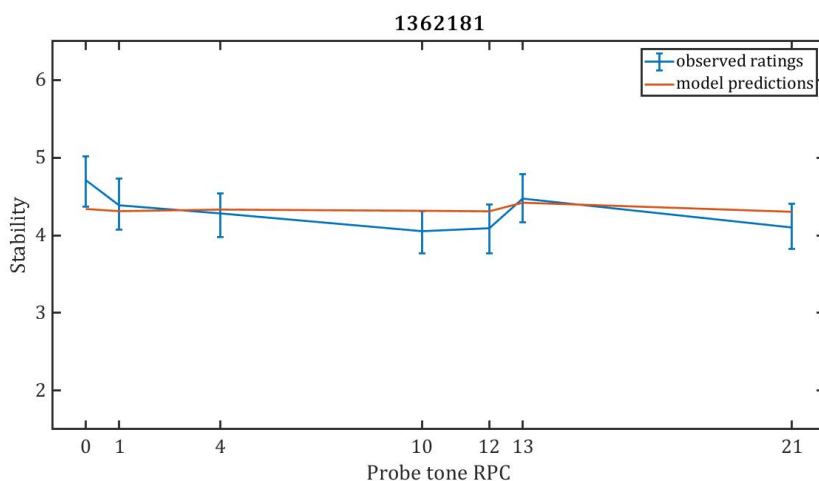


FIGURE 6.19: Average ratings for probe tones after a context of Scale 7, compared to SPCS predictions – scale-tones only.

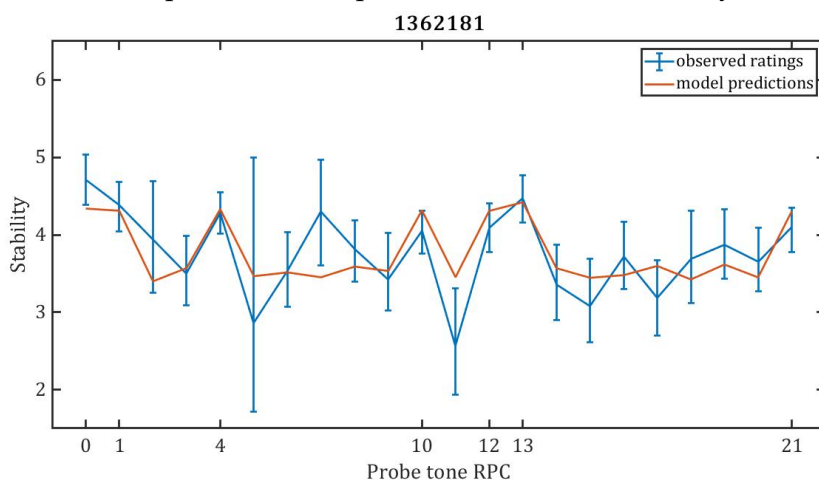


FIGURE 6.20: Average ratings for probe tones after a context of Scale 7, compared to SPCS predictions – all RPCs.

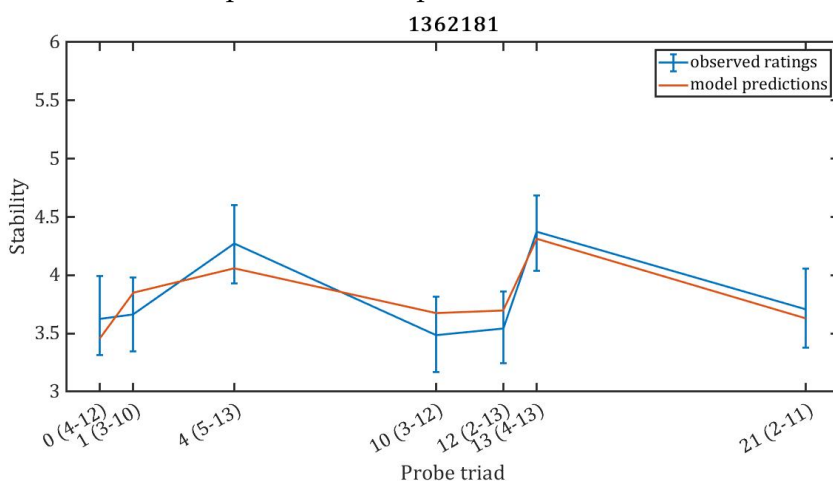


FIGURE 6.21: Average ratings for probe triads after a context of Scale 7, compared to SPCS predictions. Triad roots are labelled by RPC.

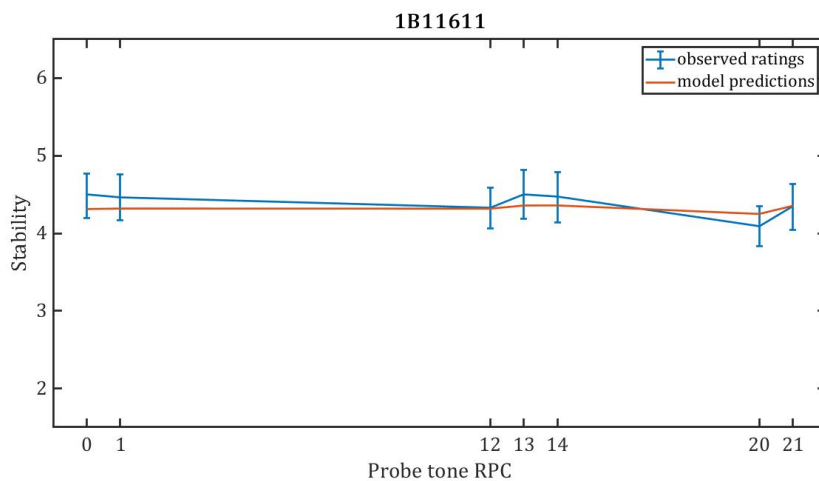


FIGURE 6.22: Average ratings for probe tones after a context of Scale 8, compared to SPCS predictions – scale-tones only.

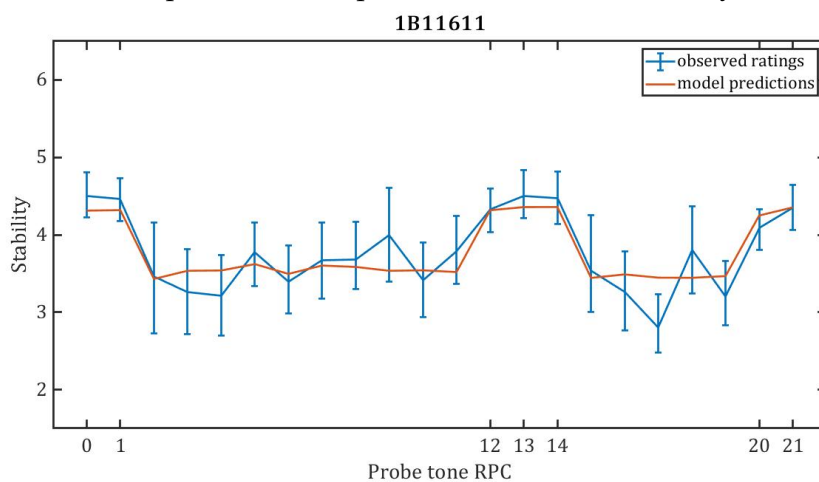


FIGURE 6.23: Average ratings for probe tones after a context of Scale 8, compared to SPCS predictions – all RPCs.

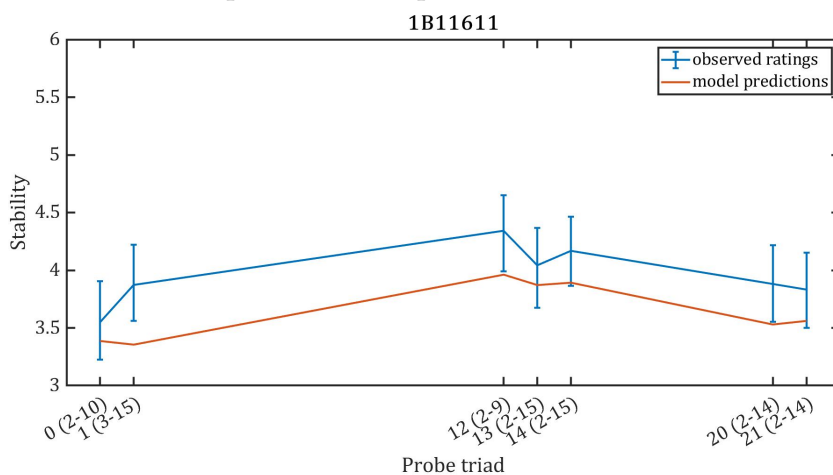


FIGURE 6.24: Average ratings for probe triads after a context of Scale 8, compared to SPCS predictions. Triad roots are labelled by RPC.

For the probe tone experiment, only for the diatonic-like scales (4324342 and 4414441) did a single RPC emerge as clearly more stable. For the probe triad experiment, however, apart from for Scale 1, when at least one triad approximating the just major triad (the most consonant triad within an octave) – namely, the scales 4324342, 4243432, and 6124612 – one of these triads was rated more stable than any other tertian triad in the scale, even when more than one triad approximates the just major triad. The highest rated (scale-tone) RPC for the probe tone experiment corresponded to the root of the highest rated triad in the probe triad experiment only for the scales 4324342 and 6124612, where a major triad is rooted on that RPC. In the probe tone ratings for the scale 6124612, RPC 21 – which is not included in the stimulus – is given the highest rating. We can assume from this that participants are sometimes confusing adjacent pitches. Given Krumhansl and Shepard (1979) and Bailes et al. (2015) find that this occurred for pitches separated by 50c, only 5.5c smaller than the interval that separates adjacent pitches of 22-TET, we are not surprised to see this occur. RPC 18 of the same scale may be another example, along with RPCs 7 and 19 from the scale 6226222. On the whole, however, it does not seem like adjacent pitches were confused with each other, which points to the possibility of additional factors at play.

We note that for the scale 6226222, for which no tonal hierarchy was found, SPCS predicts almost completely uniform stability for the scale-tone RPCs. Though for the scale-tones uniform ratings are not observed, the LOO comparison suggest that the stabilities of the RPCs across all tones are not essentially different. In all other scales, we can see that ratings of scale-tones are predicted more strongly than ratings of non-scale-tones. This is not surprising, as the scale-tones were probed around twice as much on average than the non-scale-tones, given that only a random selection of 7 of the non-scale-tone RPCs were probed. From the figures it seems that for the probe triads, the SPCS of the probe hardly contributes, with predictions almost entirely derived from the consonance of the probe triad (or, more accurately, its intrinsic stability, as measured in Experiment 3, standing in for consonance).

We can see in our analysis the emergence of effects that are not accounted for in our confirmatory analysis. Firstly, where stronger tonal hierarchies and more distinct tonics in the probe tone experiment are observed for the diatonic-like scales, we might intuit some effect of familiarity with the diatonic scale. As mentioned above, all scales bar 6226222 include at least one instance of an interval of 13 degrees – 22-TET's best perfect fifth of  $\sim 709c$  (the next best approximation of the just perfect

fifth of  $\sim 702c$  is the 12-degree interval of  $\sim 654c$ ). From the root of one of these instances, each of these scales also include an approximation to another note from the major scale, i.e., major second, major third, perfect fourth, major sixth or major seventh. For some of these scales, we observe that an RPC that approximates an additional major or perfect interval receives higher ratings even when its not a scale-tone, i.e., not being present in the stimulus. We can perhaps see that to a small degree with RPC 4 in the scale 6124612. RPCs 0, 7, 9, 13 and 20 approximate the tonic, major third, perfect fourth, perfect fifth and major seventh of a diatonic scale rooted on RPC 0. Missing only from the diatonic scale is the major second and major seventh. The major second can be approximated by RPC 4, which we perhaps see a slightly higher rating for, though it is not convincing. RPC 18 receives a much higher rating than what is predicted, though the major sixth of the diatonic scale would be better approximated by RPCs 16 or 17. We see it in the scale 1362181 with RPCs 7 and 16, where RPCs 0, 4 and 13 approximate the root and the major second and seventh of the diatonic major scale rooted on RPC 0. RPCs 7 and 16 then approximate the diatonic scale's major third and sixth, though we do not believe the ratings of RPC 16 to be high enough for consideration here. Finally, we could say it exists also in RPCs 5, and 18 of the scale 1B11611, where RPCs 1, 13 and 21 approximate the root, perfect fifth and major seventh of a diatonic scale. RPCs 5 and 18 approximate the major second and sixth of such a diatonic scale. It is perhaps worth noting that these are the least even scales, in which gaps occur to be filled in by the diatonic scale-tone approximations.

The distribution of gaps across the scale relates also to another possible effect. From the scale-tone only plots for probe tones for the two scales with the biggest gaps – 1362181 and 1B11611 – a sort of pattern emerges from the ratings that SPCS is unable to account for. The RPC at the bottom of a large step in the scale, or a 'gap', receives a higher rating than the RPC at the top of a large step.

These effects are explored in the final (third) exploratory analysis.

### 6.5.3 Confirmatory analysis

Table 6.3 shows the significant population-level effects in a Bayesian ordinal mixed effects model of the results of the probe tone experiment, as well as the associated conditional main effects and the intercepts. The effects are shown in full in Table D.1 in Appendix D.

TABLE 6.3: Model Tone significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.86	0.09	-3.03	-2.69	5373	1.00
Intercept[2]	-1.57	0.08	-1.74	-1.41	5130	1.00
Intercept[3]	-0.41	0.08	-0.57	-0.24	4995	1.00
Intercept[4]	0.24	0.08	0.07	0.40	4998	1.00
Intercept[5]	1.27	0.08	1.11	1.44	5104	1.00
Intercept[6]	2.61	0.09	2.43	2.78	5428	1.00
MusSoph	-0.13	0.08	-0.28	0.02	5453	1.00
Height	0.24	0.17	-0.08	0.56	10016	1.00
Previous	0.25	0.05	0.14	0.36	9006	1.00
Recency	0.30	0.12	0.06	0.54	14460	1.00
SPCS	0.46	0.06	0.34	0.59	9169	1.00
TrialNo	0.03	0.04	-0.04	0.10	16257	1.00
MusSoph:Recency	0.37	0.10	0.17	0.58	13741	1.00
MusSoph:SPCS	0.29	0.06	0.18	0.40	9957	1.00
Height: TrialNo	-0.20	0.10	-0.39	-0.00	10259	1.00

Significant population-level effects for Model Tone-1 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

Significant in this model as main effects are Previous (small, evid. ratio > 11999), Recency (small, evid. ratio 125.32) and SPCS (medium, evid. ratio > 11999), all positive. Additionally, Recency (medium, evid. ratio 1999) and SPCS (small, evid. ratio > 11999) both interact positively and significantly with MusSoph, where participants with higher musical sophistication respond more strongly SPCS and Recency. Finally, an interaction between Height and Trial Number (small, evid. ratio 43.28) is significant in the negative direction, where Height is more influential for earlier trials.

Table 6.4 shows the significant population-level effects in a Bayesian ordinal mixed effects model of the results of the probe triad experiment, as well as the associated conditional main effects and the intercepts. The effects are shown in full in Table D.2 in Appendix D.

TABLE 6.4: Model Triad significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.54	0.11	-2.76	-2.32	1343	1.00
Intercept[2]	-1.22	0.11	-1.43	-1.00	1303	1.01
Intercept[3]	-0.17	0.11	-0.38	0.04	1301	1.01
Intercept[4]	0.47	0.11	0.26	0.69	1307	1.01
Intercept[5]	1.62	0.11	1.40	1.84	1324	1.01
Intercept[6]	3.05	0.11	2.83	3.27	1408	1.00
MusSoph	-0.04	0.10	-0.24	0.15	1265	1.00
Previous	0.36	0.07	0.24	0.49	2827	1.00
Recency	0.05	0.06	-0.08	0.17	6359	1.00
SPCS	0.28	0.05	0.18	0.39	3584	1.00
MelCont	0.06	0.03	0.01	0.11	7975	1.00
Count	-0.01	0.05	-0.10	0.08	6349	1.00
ChordStab	0.64	0.08	0.48	0.80	2497	1.00
TrialNo	0.05	0.06	-0.06	0.16	4187	1.00
IncontTrialNo	0.07	0.03	0.01	0.13	8096	1.00
MusSoph:SPCS	0.18	0.05	0.09	0.27	3614	1.00
SPCS:TrialNo	-0.08	0.02	-0.13	-0.03	7336	1.00
Recency:IncontTrialNo	-0.11	0.04	-0.19	-0.02	8817	1.00
SPCS:IncontTrialNo	0.05	0.02	0.01	0.09	7454	1.00
Count:IncontTrialNo	-0.04	0.02	-0.08	-0.00	11062	1.00
ChordStab:IncontTrialNo	-0.05	0.02	-0.09	-0.01	8825	1.00

Significant population-level effects for Model Triad-1 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ )’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

The six intercepts represent “cutpoints” between the 7 ordinal values of ratings; the Intercept values are of a latent continuous variable; for Bayesian models in this chapter the Estimate value corresponds to the change in this latent variable that is associated with an increase of 1 standard deviation in the value of the effect for continuous variables, or with an increase from a value of 0 to a value of 1 for binary variables (Recency only, in Table 6.4). We interpret effects with a magnitude of about 0.2 to be small, those of about 0.5 to be medium, those of about 0.8 to be large.

In this model, SPCS and Previous are again significant positive effects, though the effect of SPCS (evid. ratio > 11999) is small in this model, and the effect of

Previous (evid. ratio > 11999) is of medium size. Recency drops out of significance as a main effect, significant only as a negative interaction with InContTrialNo (small, evid. ratio 125.32), where within a group of trials with the same scale as context, the effect of Recency weakens across trial number. MelCont is significant as a positive main effect for this Experiment, though rather small (evid. ratio 74), and ChordStab is positive, significant and or medium size (evid. ratio > 11999). Additional significant effects comprise interactions of SPCS, Count and ChordStab with InContTrialNo, where while the effect of SPCS increases across trial number within the trials of each context scale, the effects of Count and ChordStab decrease.

To test our first hypothesis, the Tone and Triad models were compared via LOOIC to models that differed only by the absence of SPCS for Model Tone, and by the absence of both SPCS and Triad for Model Triad. Model Tone and Model Triad were seen to significantly outperform their pairs, confirming the hypothesis.

To assess the influence of SPCS alone in Model Triad an additional comparison is made: between Model Triad and a model that differs only by the absence of SPCS. Again, Model Triad was found to perform significantly better. The same is done for Triad alone, where again Model Triad outperformed its pair. All four model comparisons are detailed in Table 6.5. We cannot yet say whether or not SPCS speaks for more in the results than simply whether or not the pitch class of the probe (Exp 4) or the number of pitch classes of the probe (Exp 5) that were also included in the context stimulus. Accordingly, Models 6.1 and 6.22 were compared via LOOIC to models in which SPCS is replaced by ‘ScaleTone’, which accounts for this simple measure.

For both experiments, the models including ScaleTone significantly underperformed those including SPCS, as shown in Table 6.5.

TABLE 6.5: LOOIC comparisons for Bayesian ordinal mixed effects models

Model compared to Model Tone	Model – Model Tone LOOIC	SE	Signif
– SPCS	1139.6	79.6	yes
– SPCS + ScaleTone	135.2	18.8	yes
Model compared to Model Triad	Model - Model Triad LOOIC	SE	Signif
– Triad – SPCS	2251.8	111.6	yes
– Triad	1972.4	103.8	yes
– SPCS	352.6	47.4	yes
– SPCS + ScaleTone	27.2	7.4	yes



### 6.5.4 Exploratory analyses

As detailed in Chapter 5, our scales were chosen such that they best represent the entire distribution of possible heptatonic scales of 22-TET in a 13-dimensional tonal scale feature space. Accordingly, if responses differ for different scales, we may hope to gain some insight into the perceptual reality of these features. In order to gain such insight, we ran two additional analyses.

1. A categorical effect of Scale was added to the models as an interaction with SPCS, Recency, Height etc. in the same way the MusSoph and TrialNumber are included in the models.
2. Instead of a categorical effect of scale, 13 effects were added in the same way, where each effect is the value of each tonal scale feature for the scale used in stimulus for the trial, as detailed in Chapter 5.

Table 6.6 shows the significant effects and their associated marginal effects in the first of these exploratory models for the probe tone experiment. See Table D.3 in Appendix D for the complete list of effects.

We can see from this that the effect of SPCS, Recency and Previous differs between scales. The second exploratory analysis should lead us towards an explanation of what features of the scales lead to this difference in response.

TABLE 6.6: Exploratory Model 6.1 population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.01	0.12	-3.24	-2.78	3467	1.00
Intercept[2]	-1.65	0.11	-1.87	-1.43	3279	1.00
Intercept[3]	-0.42	0.11	-0.63	-0.20	3257	1.00
Intercept[4]	0.26	0.11	0.05	0.48	3277	1.00
Intercept[5]	1.35	0.11	1.14	1.57	3297	1.00
Intercept[6]	2.75	0.11	2.53	2.97	3463	1.00
MusSoph	-0.11	0.08	-0.28	0.05	2900	1.00
Previous	0.17	0.07	0.03	0.31	4843	1.00
Recency	0.38	0.20	-0.01	0.76	5101	1.00
SPCS	0.46	0.08	0.31	0.62	3554	1.00
Scale2	-0.10	0.11	-0.31	0.11	7890	1.00
Scale3	0.13	0.10	-0.07	0.33	7496	1.00
Scale4	-0.06	0.13	-0.32	0.19	6990	1.00
Scale5	0.08	0.12	-0.16	0.32	7459	1.00
Scale6	-0.01	0.10	-0.20	0.18	7309	1.00
Scale7	-0.07	0.12	-0.31	0.16	6994	1.00
Scale8	0.01	0.10	-0.19	0.20	7746	1.00
MusSoph:Recency	0.40	0.11	0.18	0.62	8644	1.00
MusSoph:SPCS	0.29	0.06	0.17	0.41	3986	1.00
Previous:Scale2	0.16	0.08	0.01	0.32	6842	1.00
Previous:Scale3	-0.01	0.10	-0.19	0.19	7468	1.00
Previous:Scale4	-0.08	0.08	-0.23	0.07	7388	1.00
Previous:Scale5	0.12	0.08	-0.03	0.28	6853	1.00
Previous:Scale6	0.02	0.07	-0.13	0.16	7603	1.00
Previous:Scale7	0.11	0.09	-0.05	0.28	6454	1.00
Previous:Scale8	0.03	0.08	-0.12	0.19	7978	1.00
Recency:Scale2	-0.53	0.27	-1.05	-0.00	7246	1.00
Recency:Scale3	0.12	0.27	-0.40	0.65	6816	1.00
Recency:Scale4	-0.11	0.27	-0.62	0.42	6591	1.00
Recency:Scale5	0.43	0.30	-0.15	1.04	7062	1.00
Recency:Scale6	-0.22	0.26	-0.74	0.28	6664	1.00
Recency:Scale7	-0.07	0.28	-0.62	0.47	7022	1.00
Recency:Scale8	-0.04	0.28	-0.59	0.51	6431	1.00
SPCS:Scale2	0.30	0.08	0.14	0.46	5976	1.00
SPCS:Scale3	-0.01	0.08	-0.17	0.15	6116	1.00
SPCS:Scale4	0.07	0.08	-0.09	0.23	5967	1.00
SPCS:Scale5	-0.14	0.09	-0.32	0.05	6672	1.00
SPCS:Scale6	-0.09	0.08	-0.24	0.06	6085	1.00
SPCS:Scale7	-0.03	0.09	-0.20	0.14	6760	1.00
SPCS:Scale8	0.11	0.09	-0.06	0.30	6178	1.00

Significant population-level effects for Exploratory Model 6.1 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

Table 6.7 shows the evidence ratios of each significant effect in the model.

TABLE 6.7: Evidence ratios for all significant effects of Exploratory Model 6.1

Hypothesis	evidence ratio
SPCS > 0	> 11999
Previous > 0	129.43
Recency > 0	36.27
MusSoph:SPCS > 0	> 11999
MusSoph:Recency > 0	2999
Scale3 > Scale2	47
Previous:Scale2 > Previous:Scale1	54.56
Previous:Scale2 > Previous:Scale4	599
Previous:Scale5 > Previous:Scale4	135.36
Previous:Scale7 > Previous:Scale4	56.97
Previous:Scale2 > Previous:Scale6	29.93
SPCS:Scale2 > SPCS:Scale1	> 11999
SPCS:Scale2 > SPCS:Scale3	> 11999
SPCS:Scale2 > SPCS:Scale4	254.32
SPCS:Scale2 > SPCS:Scale5	> 11999
SPCS:Scale2 > SPCS:Scale6	> 11999
SPCS:Scale2 > SPCS:Scale7	11999
SPCS:Scale2 > SPCS:Scale8	32.8
SPCS:Scale4 > SPCS:Scale5	50.5
SPCS:Scale4 > SPCS:Scale6	30.91
SPCS:Scale8 > SPCS:Scale5	98.17
SPCS:Scale8 > SPCS:Scale6	64.93
SPCS:Scale2 > SPCS:Scale1	> 11999
Recency:Scale1 > Recency:Scale2	39.4
Recency:Scale5 > Recency:Scale2	856.14
Recency:Scale7 > Recency:Scale2	74
Recency:Scale8 > Recency:Scale2	20.7
Recency:Scale5 > Recency:Scale4	20.51
Recency:Scale5 > Recency:Scale7	47.58

In a second exploratory analysis Scale was replaced by the 13 scale features that were used in Chapter 5 in order to arrive at the selection of scales used in this experiment. As a reminder, grouped into six types of scale features, the 13 that the reduction leads to are as follows:

- Redundancy:
  1. maximum variety
  2. well-formedness
  3. pairwise well-formedness
  4. trichord entropy
  5. pentachord entropy
- Coherence and evenness:
  6. Lumma stability
  7. Lumma impropriety
- R-ad entropy:
  8. triad entropy
- Generator complexity:
  9. scalar Graham complexity
- Consonance:
  10. min consonance
  11. max consonance
  12. median consonance
- Tetrachordality
  13. tetrachordality

Considering that the full model was too large to run within a reasonable time span with the addition of these 13 features, all interacting with MusSoph, SPCS, Recency, etc., we removed the effects from the model that were not significant in the first exploratory model, leading to:

```
brm(Ratings ~ 1 + (MusSoph + SGC + TriadEnt + MaxVar + WF + PWF + Tri-
chordEnt + PentachordEnt + LummaStblty + LummaImprty + MaxCons + MinCons
+ MedCons + Tetrachrdlty)*(Previous + Recency + SPCS) + (1 + (SGC + TriadEnt
+ MaxVar + WF + PWF + TrichordEnt + PentachordEnt + LummaStblty + Lum-
maImprty + MaxCons + MinCons + MedCons + Tetrachrdlty)*(Previous + Recency
+ SPCS) | Number)
```

Only MusSoph:SPCS (positive, small, evid. ratio) and MusSoph:Recency (negative, small, evid. ratio) were significant in the model, whose effect sizes are shown in Table D.4 in Appendix D. None of our 13 scale features are significant in the model. Where we know from the results of the first exploratory model that significantly differences in ratings do exist between some scales, this analysis suggests that our features that we had suggested may relate to perceived tonality in scales do not in fact affect stability ratings of RPCs in our scales. We note, however, that we are testing in this model for the effect of 13 scales features on only 8 scales. It would be wise in the future to run an experiment testing for the effects of a single scale feature upon the perception of stability in the RPCs and triads of microtonal scales.

One final exploratory model was run, including predictors of the effects observed in the descriptive model for the probe tone experiment.

We will first define DiatSim and ChromSim as the similarity between the context

scale and the 12-TET diatonic and chromatic scales, respectively, considering, unlike for SPCS, only the fundamental pitches.

Additionally, to account for the higher ratings given to RPCs that correspond reasonably closely to RPCs of a 12-TET diatonic scale, we model an effect of DiatBoost, which boosts the predicted ratings for RPCs that approximate RPCs of the 12-TET diatonic scale. To achieve this, the context scale is represented by a vector of 1200 entries (corresponding to 1200 cents in an octave). Seven of these 1200 entries carry the value '1' – those corresponding to the cents values of the seven RPCs of the context scale in 22-TET. The 12-TET diatonic scale is represented first as a list of the number of occurrences in the scale of each possible interval of 12-TET (up to an octave), i.e., (0 2 5 4 3 6 2 6 3 4 5 2) – 0 unisons, 2 semitones, 5 tones, etc. This vector is then represented by an expectation vector of 1200 entries in the same way as for the context scale above, with the entries corresponding to cents values of (100 200 300 400 500 600 700 800 900 1000 1100) carrying the values (2 5 4 3 6 2 6 3 4 5 2). The value of DiatBoost for an RPC of the context scale is the value of the convolution of these two vectors at the entries corresponding to the cents values of the RPC, after Gaussian smoothing is applied as for SPCS (see Appendix A for the mathematical definition of this and of expectation vectors).

Finally, ModeHeight is added to account for the observed effect of RPCs receiving higher ratings when the mode of the scale of which they would be the tonic has higher average pitch height. The ModeHeight value of each RPC is calculated as the average pitch height of the mode of the scale that begins on the RPC. We will explain how to calculate a mode's average pitch height – its *mode height* – using the major or Ionian mode of the Superpythagorean diatonic scale in 22-TET (introduced in Chapter 5 originally as the WF diatonic scale of 22-TET in Section 5.3.2, and Scale 4 here) – 4414441 — as an example:

1. Express the mode as degrees above 0: 4, 8, 9, 13, 17, 21, 22.
2. Remove the top note to centre the mode: 4, 8, 12, 13, 17, 21.
3. Calculate the deviation of each of the remaining 6 notes from the centre of the octave, which for a mode of a scale of  $N$ -TET is equal to  $N/2$  for (in this case, 11): -7, -3, -2, 2, 6, 10.
4. Take the sum of the values. In this case, we arrive at a mode height of 6 degrees.

The modes for Scale 4, 22-TET's Superpythagorean diatonic scale are as such:

- Mode 9: 4441441
- Mode 6: 4414441
- Mode 3: 4414414
- Mode 0: 4144414
- Mode -3: 4144144
- Mode -6: 1444144
- Mode -9: 1441444

Though not integral to this project, it may be worth mentioning here that for WF scales the heights of the modes are multiples of the scale's *chroma*: The difference between the large and small steps. For WF scales of size  $S \geq 5$ , the heights of the modes are

$$-c(S-1)/2, -c(S-3)/2, \dots, 0, \dots, c(S-3)/2, c(S-1)/2,$$

where  $c$  is the scale's *chroma* in degrees of the ET to which the scale is tuned.

Recalling its examination in Section 5.3.6, this scale's modes and their average pitch heights are as thus:

The mode beginning on RPC 0, as defined based upon our results, is the major mode, Mode 6, which comprises RPCs 0, 4, 8, 9, 13, 17, and 21. Since Mode 6 begins on RPC 0, RPC 0 has ModeHeight 6; Mode 0 begins on RPC 4, so RPC 4 has ModeHeight 0; Mode -6 begins on RPC 8, so RPC 8 has ModeHeight -6; Mode 9 begins on RPC 9; so RPC 9 has ModeHeight 9; etc.

DiatBoost and ModeHeight are included in the same way as SPCS and Recency, etc. in the model as effects that describe strategies for completing the task, and DiatSim and ChromSim are included in the same way as MusSoph and TrialNo etc. as effects that impact the strength of the effects that describe the strategies.

The resulting model is as follows:

$$\text{brm}(\text{Ratings} \sim 1 + (\text{MusSoph} + \text{TrialNo} + \text{IncontTrialNo} + \text{I}(\text{TrialNo}^2) + \text{ChromSim} + \text{DiatSim}) * (\text{Height} + \text{RelHeight} + \text{I}(\text{Height}^2) + \text{I}(\text{RelHeight}^2)) + \text{Primacy} + \text{Previous} + \text{Recency} + \text{SPCS} + \text{MelCont} + \text{Count} + \text{ModeHeight} + \text{DiatBoost}) + (1 + (\text{TrialNo} + \text{IncontTrialNo} + \text{I}(\text{TrialNo}^2) + \text{ChromSim} + \text{DiatSim}) * (\text{Height} + \text{RelHeight} + \text{I}(\text{Height}^2) + \text{I}(\text{RelHeight}^2)) + \text{Primacy} + \text{Previous} + \text{Recency} + \text{SPCS} + \text{MelCont} + \text{Count} + \text{ModeHeight} + \text{DiatBoost}) | \text{Number})$$

Table 6.8 below displays the effects significant in this model, along with their conditional effects and the intercepts. The model's effects are shown in full in Table D.5 in Appendix D.

TABLE 6.8: Exploratory Model 6.3 population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.92	0.09	-3.10	-2.74	3069	1.00
Intercept[2]	-1.61	0.09	-1.78	-1.43	2904	1.00
Intercept[3]	-0.41	0.09	-0.58	-0.24	2838	1.00
Intercept[4]	0.25	0.09	0.08	0.42	2836	1.00
Intercept[5]	1.32	0.09	1.15	1.49	2927	1.00
Intercept[6]	2.70	0.09	2.52	2.88	3101	1.00
MusSoph	-0.13	0.08	-0.29	0.03	2586	1.00
TrialNo	0.03	0.04	-0.05	0.10	10652	1.00
ChromSim	-0.01	0.05	-0.11	0.09	8150	1.00
DiatSim	-0.02	0.06	-0.14	0.09	7195	1.00
Previous	0.24	0.05	0.14	0.34	5745	1.00
Recency	0.32	0.13	0.08	0.57	10074	1.00
SPCS	0.44	0.06	0.32	0.56	5276	1.00
ModeHeight	0.03	0.03	-0.03	0.10	11508	1.00
DiatBoost	0.04	0.04	-0.03	0.11	10233	1.00
MusSoph:Recency	0.38	0.11	0.17	0.60	9721	1.00
MusSoph:SPCS	0.26	0.05	0.16	0.37	5682	1.00
MusSoph:ModeHeight	0.08	0.02	0.04	0.13	13014	1.00
MusSoph:DiatBoost	0.08	0.03	0.03	0.14	9209	1.00
TrialNo:DiatBoost	0.05	0.02	0.01	0.09	14452	1.00
ChromSim:Previous	0.11	0.04	0.04	0.18	11548	1.00
DiatSim:Previous	-0.10	0.04	-0.18	-0.03	10868	1.00
DiatSim:SPCS	0.11	0.05	0.02	0.20	9361	1.00

Significant population-level effects for Exploratory Model 6.3 along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

Apart from the interaction between Height and Trial No, all effects significant in Model Tone (see Table 6.3) are significant also in this model (i.e., the 95% CIs around their estimates do not cross zero). Additionally, we see positive interactions of MusSoph with both ModeHeight and DiatBoost, suggesting that these alternative strategies used in ratings are adopted to a larger extent by musicians, similar to SPCS and Recency, though unlike Recency and SPCS the conditional main effects

of ModeHeight and DiatBoost are not significant. We also see as significant, a positive interaction between TrialNo and DiatBoost, suggesting that DiatBoost affects ratings to a larger extent later in the experiment; a positive interaction between Previous and ChromSim, and a negative interaction between Previous and DiatSim, suggesting that when the context scale was more similar to the 12-TET chromatic scale and less similar to the 12-TET diatonic scale, participants relied more on Previous to inform their ratings; and finally a positive interaction between SPCS and DiatSim, suggesting that SPCS has a stronger influence on participants' stability ratings for scales more similar to the 12-TET diatonic, reflecting what we saw in previous analyses.

This model was compared via LOOIC to Model Tone (Table 6.3), and found to significantly outperform it, with a difference in LOOIC of 150.6 for a SE of 43.2. We suggest that in the future another experiment should be run in order confirm the superiority of this exploratory model.

## 6.6 Discussion and Conclusion

We have demonstrated in these experiments that SPCS is able to predict the perceived stability of tones and triads in novel scales. For the probe tone experiment, tonal hierarchies were observed in all but one of eight novel microtonal heptatonic scales chosen to represent the entire distribution of possible heptatonic scales of 22-TET according to a set of 13 scale features. We noted this scale to be the only scale tested that lacks any instances of 22-TET's best perfect fifth. Noting that for Experiment 1 the only scale from which a tonal hierarchy was not uncovered was the only scale which did not include a perfect fifth, we are tempted to suggest that the presence of a perfect fifth may be necessary for a scale to be able to support tonal hierarchy. In light of what's understood in music theory about the importance of the perfect fifth in tonal harmony, and the frequency with which it occurs in music (Huron, 1991, 1994), this is not surprising. Recalling, however, Loui's findings (Loui et al., 2010) that musical grammars can be learned rapidly in the Bohlen-Pierce scale, which does not contain a decent approximation of a perfect fifth, we clarify that though perhaps statistical learning can induce a hierarchy in any arbitrary scale, only for a scale with a perfect fifth does a tonal hierarchy emerge intrinsically (given the differences in the context stimulus across Loui's and our experiments). We might



make a similar suggestion concerning the probe triad experiment – that an approximation of a major triad is necessary for tonal-harmonic hierarchy – given that we observed a tonal-harmonic hierarchy for all scales with approximations to major triads. Inspection of the plots of the descriptive models shows that for the probe tone experiment, where hierarchies were observed, the highest ratings were given to RPCs over which lie a perfect fifth. This is not at odds with SPCS, as pitch classes separated by a perfect fifth have very high SPCS, given that low integer frequency ratios correspond to high coincidences of overtones.

We noted that for the probe tone experiment, these hierarchies can be predicted using SPCS. In Bayesian ordinal mixed effect regression models of stability ratings, SPCS was found to be significant as a conditional main effect as well as as an interaction with MusSoph, and removing SPCS from the models significantly decreased their performance. Exploratory analyses of the probe tone experiment suggest that the strength of predictors varies for different scales, though we were unable to find significant effects of any of the 13 scale features. Perhaps though mathematically appealing they are not perceptually relevant, however if we wish to test for the effect of any scale features, we should ideally run an experiment that controls only a few of the 13 features and far more than 8 scales, as this experiment was not ideal for such a test. A final exploratory analysis however does support our post-hoc hypothesis that the similarity of the context scale to the 12-TET diatonic scale does affect the strength of some predictors. The final exploratory analysis also supports our hypothesis that the effects DiatBoost and ModeHeight significantly impact ratings, where these effects were significant as positive interactions with MusSoph (along with SPCS and Recency). The model run in this final analysis was compared to Model Tone and found to significantly outperform it, further supporting the validity and utility of these effects. Further research can be done in order to test this exploratory model.

## Chapter 7

# General Discussion and Conclusions

In this thesis we uncovered tonal hierarchies in a number of familiar, unfamiliar and novel scales. These hierarchies were able to emerge from the pitches of the scales alone, with the influence of other musical elements minimized. We were able to predict the perceived stabilities of RPCs and triads in these scales using SPCS, as well as using predictors built from the prevalence of RPCs and triads in the scales from an appropriate corpus, finding clear evidence for psychoacoustic and statistical learning influences on the perception of harmonic tonality. We also systematically explored a set of over 7000 novel microtonal scales in terms of many mathematically appealing features. This research should pave the way for further exploration of harmonic tonality in microtonal scales.

Noting that in previous literature – including Krumhansl (2001) – ratings of goodness-of-fit have been equated with perceived musical stability a pair of experiments were run in which participants rated either fit or stability. We found that though differing for specific RPCs and triads in familiar scales (Experiment 1) ratings of fit were overall equivalent to ratings of stability, given our stimulus. The RPCs for which significant differences were found aligned with music-theoretical discourse. This suggests the existence of a learned response, in addition to the psychoacoustic response our hypotheses concerned. Concerning Experiment 1 again, the ability for the statistical prevalence of RPCs and triads in a rock-pop corpus to predict our results strengthens the argument for the effect of statistical learning on responses. In Experiment 2, as well as in Experiment 1, significant effects of SPCS (a measure of the psychoacoustic response we tested for) and its interaction with musical sophistication were found, and models of the data were in most cases significantly weakened by its removal. Since the scales used in Experiment 2, though less familiar overall, were not novel, this interaction suggests either that long-term statistical learning of scales is required to differentiate between RPCs or triads in the

scales, or that musical training improves audition skills, which involve the ability to differentiate RPCs and triads in scalic contexts.

After Experiment 3 recorded the intrinsic stability of all triads in 22-TET and a distributional analysis led to the selection of scales of 22-TET, Experiments 4 and 5 tested the perceived stability of RPCs and tertian triads in these novel scales. SPCS and its interaction with musical sophistication were again found to be significant in the models of the results, and the removal of SPCS from the models significantly weakened them. Since these scales were novel at least to our listeners (some of the scales have been used in compositions by the author, and other contemporary composers; (“22edo”, 2020)), we may interpret from these results that the perceived stability of RPCs and triads in scales is due, at least in part, to a psychoacoustic process concerning the spectral content of the stimulus, and may be predicted by SPCS. Further, the strength of this mechanism is increased through musical training.

We should note, however, that exploratory analyses suggest that effects of familiarity upon ratings are still present for these novel scales, weighted by their similarity to the familiar 12-TET diatonic and chromatic scales. We recall the familiarity effects observed in our initial analysis: the similarity of the scale to 12-TET and to the 12-TET diatonic scale, and the apparent boost given to the perceived stabilities of RPCs that line up with RPCs of the 12-TET diatonic mode most similar to the context scale. The analysis exploring these effects could be confirmed with future experiments.

Though our first exploratory analysis for Experiment 4 revealed significant interactions with a categorical effect of context scale upon stability ratings, in our second exploratory analysis, which included our 13 scale features used in the distributional analysis instead of a categorical scale effect, none of these features came through as significant effects. To test for the effect of these scale features it would be wise in future work to run an experiment testing for the effects of a single scale feature upon the perception of stability of the RPCs and triads of microtonal scales, instead of testing for the effects of 13 features at once. It is still possible of course that these features, though mathematically appealing, do not reflect a perceptually reality. If this were true for all features though, then what would be the features that do influence perception? We expect that this more appropriate analysis will find some of our features to be pertinent.

A simple model comparison was run to test for the existence of a tonal hierarchy in triads and for RPCs in each scale in Experiments 1, 2, 4 and 5: A model with

only the intercept as a predictor (a null model) was compared to a model with only the identity of the probe as a predictor. When the null model was found to be the weaker of the pair to a significant degree, we suggested that a tonal hierarchy was present. Hierarchies were uncovered in this way for all scales of Experiment 1 and 2, for RPCs and triads, apart from for the octatonic scale. We noted, however, that only two hierarchical levels could be seen in the hexatonic scale – the scale-tones and the non-scale-tones. Perhaps two hierarchical levels within the scale-tones, with the non-scale tones representing a third, lower level, should be required at the least. In future research we aim to test specifically for the existence of tonal hierarchies in familiar and novel scales using a more sophisticated analysis that tests for at least 3 hierarchical levels. We might also seek to test whether or not the top hierarchical level requires membership by only a single RPC or triad for the employment of tonal harmony in a scale, given that a unique tonic is employed in much tonal-harmonic music. We should consider the role in other musical features in the establishment of a tonic. We could also consider running an experiment testing how a keyboard improviser uses our set of novel scales statistically, and how that relates to the perceived hierarchies.

Hierarchies were uncovered for all scales of Experiments 4 and 5 that contain the tuning's best approximation of a JI perfect fifth. We noted that this was true also of our 12-TET scales, tested over Experiments 1 and 2, for which the octatonic scale was the only scale lacking 12-TET's perfect fifth. We hypothesize that the existence of a decent approximation of a JI perfect fifth in a scale is required for the formation of a tonal hierarchy. We aim to test for this hypothesis in the future research introduced immediately above.

For Experiments 1 and 2 the consonance of triads was captured by a categorical effect of triad; the perceived intrinsic stability of triads as recorded in Experiment 3 was used in Experiments 4 and 5. We intend in future research to complete the analysis concerning the second hypothesis of Experiment 3, namely that the intrinsic stability of triads may be predicted by their harmonicity, spectral entropy, sensory dissonance, harmonic entropy and familiarity. We can see already where such an analysis could be illuminating in the analysis of Experiment 3, particularly looking back to Table 4.2. We see the influence in the ratings of the simplicity of frequency ratios approximated by both the constituent intervals, relating to sensory dissonance, and of the triad as a whole, relating to harmonic entropy, harmonicity, and spectral entropy, as well as to the pull of the familiarity of the 12-TET major triad. A model

of intrinsic stability of triads built from the results of these experiments can be used in a later analysis of the data from Experiments 5 instead of the values recorded in Experiment 3, and an experiment could be run in the future to test the model against a different selection of novel triads.

Further, we may look into possibilities for music production that may arise from this work. This thesis represents the first part of a larger project, the second part of which will be the facilitation of the composition of tonal-harmonic music in novel, microtonal scales. Given that tonal hierarchies were observed for most of our microtonal scales, and these scales are a representation of all possible 7-note scales of 22-TET, we should be able to write tonal-harmonic music in most 7-note scales in 22-TET. We hypothesize that this should generalize into other tunings systems, and into scales of different numbers of notes, and further work should test this hypothesis.

We define *stability profiles* as the perceived stability of RPCs, chords, or of any musical object across time. These profiles could be studied for tonal-harmonic music. Then, using our model for perceived stability of triads and RPCs in arbitrary scales, music can be composed following common or particular stability profiles. This may maximize the possible emotional affect of music in novel scales. Response to such music can be recorded in a final experiment for the project. It should be possible then to produce some sort of guide for the composition of tonal-harmonic music in arbitrary scales that produces a desired response in listeners, which would greatly increase the scope for expression through music.

# Bibliography

- 12edo [Accessed: 2021-01-06]. (2021). <https://en.xen.wiki/w/12edo>
- 22edo [Accessed: 2020-02-11]. (2020). <https://en.xen.wiki/w/22edo>
- Agresti, A. (2010). *Analysis of Ordinal Categorical Data* (Vol. 656). New Jersey, NY, John Wiley & Sons.
- Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 19(4), 1483–1536.
- Amiot, E. (2009). Discrete fourier transform and bach's good temperament. *Music Theory Online*, 15(2). [www.mtosmt.org/issues/mto.09.15.2/mto.09.15.2.amiot.html](http://www.mtosmt.org/issues/mto.09.15.2/mto.09.15.2.amiot.html)
- Bailes, F., Dean, R. T., & Broughton, M. C. (2015). How different are our perceptions of equal-tempered and microtonal intervals? a behavioural and EEG survey. *PloS one*, 10(8).
- Balzano, G. J. (1982). The pitch set as a level of description for studying musical pitch perception. In *Music, Mind, and Brain* (pp. 321–351). Springer.
- Barbour, J. M. (2004). *Tuning and Temperament: A Historical Survey*. Courier Corporation.
- Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3), 255–278.
- Berlyne, D. E., McDonnell, P., Nicki, R. M., & Parham, L. C. C. (1967). Effects of auditory pitch and complexity on EEG desynchronization and on verbally expressed judgments. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, 21(4), 346.
- Bigand, E. (1997). Perceiving musical stability: The effect of tonal structure, rhythm, and musical expertise. *Journal of Experimental Psychology: Human Perception and Performance*, 23(3), 808.
- Bigand, E., & Parncutt, R. (1999). Perceiving musical tension in long chord sequences. *Psychological Research*, 62(4), 237–254.

- Bigand, E., Parncutt, R., & Lerdahl, F. (1996). Perception of musical tension in short chord sequences: The influence of harmonic function, sensory dissonance, horizontal motion, and musical training. *Perception & Psychophysics*, 58(1), 125–141.
- Bigand, E., & Pineau, M. (1997). Global context effects on musical expectancy. *Perception & Psychophysics*, 59(7), 1098–1107.
- Blackwood, E. (1980). *Twelve Microtonal Etudes: For Electronic Music Media*. E. Blackwood.
- Blackwood, E. (1985). *The Structure of Recognizable Diatonic Tunings*. Princeton University Press. <http://www.jstor.org/stable/j.ctt7ztvms>
- Bosanquet, R. H. M. (1878). On the hindoo division of the octave, with some addition to the theory of systems of the higher orders. *Proceedings of the Royal Society of London*, 26(179-184), 372–384.
- Branca, G. (1993). *Symphony no. 3:(Gloria): Music for the First 127 Interoals of the Harmonic Series*. Atavistic.
- Bucht, S., & Huovinen, E. (2004). Perceived consonance of harmonic intervals in 19-tone equal temperament, In *Proceedings of the Conference on Interdisciplinary Musicology (CIM04)*. Citeseer.
- Budge, H. (1943). *A Study of Chord Frequencies: Based on the Music of Representative Composers of the Eighteenth and Nineteenth Centuries*. Teachers College, Columbia University.
- Bürkner, P.-C. (2017). brms: An R package for Bayesian multilevel model using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>
- Bürkner, P.-C. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, 10(1), 395–411. <https://doi.org/10.32614/RJ-2018-017>
- Carey, N. (2002). On coherence and sameness, and the evaluation of scale candidacy claims. *Journal of Music Theory*, 46(1/2), 1–56.
- Carey, N. (2007). Coherence and sameness in well-formed and pairwise well-formed scales. *Journal of Mathematics and Music*, 1(2), 79–98.
- Carey, N., & Clampitt, D. (1989). Aspects of well-formed scales. *Music Theory Spectrum*, 11(2), 187–206.
- Carlos, W. (2000). *Beauty in the Beast*. East Side Digital.
- Castellano, M. A., Bharucha, J. J., & Krumhansl, C. L. (1984). Tonal hierarchies in the music of North India. *Journal of Experimental Psychology: General*, 113(3), 394.

- Chertoff, M. E., & Hecox, K. E. (1990). Auditory nonlinearities measured with auditory-evoked potentials. *The Journal of the Acoustical Society of America*, 87(3), 1248–1254.
- Chertoff, M. E., Hecox, K. E., & Goldstein, R. (1992). Auditory distortion products measured with averaged auditory evoked potentials. *Journal of Speech, Language, and Hearing Research*, 35(1), 157–166.
- Chowning, J. M. (2012). *Stria*. CCRMA.
- Clampitt, D. (1998). Pairwise well-formed scales: Structural and transformational properties.
- Clampitt, D. (2007). Mathematical and musical properties of pairwise well-formed scales, In *International Conference on Mathematics and Computation in Music*. Springer.
- Clough, J., & Douthett, J. (1991). Maximally even sets. *Journal of Music Theory*, 35(1/2), 93–173.
- Clough, J., Douthett, J., Ramanathan, N., & Rowell, L. (1993). Early indian heptatonic scales and recent diatonic theory. *Music Theory Spectrum*, 15(1), 36–58.
- Clough, J., Engebretsen, N., & Kochavi, J. (1999). Scales, sets, and interval cycles: A taxonomy. *Music Theory Spectrum*, 21(1), 74–104.
- Clough, J., & Myerson, G. (1985). Variety and multiplicity in diatonic systems. *Journal of Music Theory*, 29(2), 249–270.
- Consistent [Accessed: 2020-02-11]. (2019). <https://en.xen.wiki/w/Consistent>
- Cook, N. D. (2009). Harmony perception: Harmoniousness is more than the sum of interval consonance. *Music Perception: An Interdisciplinary Journal*, 27(1), 25–42.
- Cook, N. D., & Fujisawa, T. X. (2006). The psychophysics of harmony perception: Harmony is a three-tone phenomenon.
- Cook, N. D., Fujisawa, T. X., & Konaka, H. (2007). Why not study polytonal psychophysics?
- Cook, N. D., Fujisawa, T., & Takami, K. (2004). A psychophysical model of harmony perception, In *Proceedings of the 8th International Conference on Music Perception and Cognition (ICMPC8)*.
- Corrigall, K. A., & Trainor, L. J. (2009). Effects of musical training on key and harmony perception. *Annals of the New York Academy of Sciences*, 1169(1), 164–168.



- Cousineau, M., McDermott, J. H., & Peretz, I. (2012). The basis of musical consonance as revealed by congenital amusia. *Proceedings of the National Academy of Sciences*, *109*(48), 19858–19863.
- Craton, L. G., Juergens, D. S., Michalak, H. R., & Poirier, C. R. (2016). Roll over Beethoven? an initial investigation of listeners' perception of chords used in rock music. *Music Perception: An Interdisciplinary Journal*, *33*(3), 332–343.
- Craton, L. G., Lee, J. H. J., & Krahe, P. M. (2019). It's only rock 'n roll (but I like it): Chord perception and rock's liberal harmonic palette. *Musicae Scientiae*, 1029864919845023.
- Cuddy, L. L. (1993). Melody comprehension and tonal structure. In T. J. Tighe & W. J. Dowling (Eds.), *Psychology and music: The understanding of melody and rhythm* (pp. 19–38). Lawrence Erlbaum Associates, Inc.
- Cuddy, L. L., & Badertscher, B. (1987). Recovery of the tonal hierarchy: Some comparisons across age and levels of musical experience. *Perception & Psychophysics*, *41*(6), 609–620.
- Cuddy, L. L., & Smith, N. A. (2000). Perception of tonal pitch space and tonal tension. *Musiology and Sister Disciplines*, 47–59.
- Dahlhaus, C., Anderson, J., Wilson, C., Cohn, R., & Hyer, B. (1980). Harmony. *The New Grove Dictionary of Music and Musicians*, *10*, 858–877.
- Daniélou, A. (1995). *Music and the Power of Sound: The Influence of Tuning and Interval on Consciousness*. Rochester, Vermont, Inner Traditions.
- De Cheveigne, A. (2005). Pitch perception models. In *Pitch* (pp. 169–233). Springer.
- De Clercq, T., & Temperley, D. (2011). A corpus analysis of rock harmony. *Popular Music*, *30*(1), 47–70.
- Dean, R. T., Bailes, F., & Brennan, D. (2008). Microtonality, the octave, and novel tunings for affective music. *Music of the Spirit: Asian-Pacific Musical Identity*, 128–139.
- Dean, R. T., Milne, A. J., & Bailes, F. (2019). Spectral pitch similarity is a predictor of perceived change in sound- as well as note-based music. *Music & Science*, *2*.
- Deutsch, D. (1975). The organization of short-term memory for a single acoustic attribute. *Short-term Memory*, 107–151.
- Deutsch, D. (1980). The processing of structured and unstructured tonal sequences. *Perception & Psychophysics*, *28*(5), 381–389.
- Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, *25*(1), 207–218.

- Dyson, B. J., & Quinlan, P. T. (2010). Decomposing the Garner interference paradigm: Evidence for dissociations between macrolevel and microlevel performance. *Attention, Perception, & Psychophysics*, 72(6), 1676–1691.
- Eberlein, R. (1994). *Die Entstehung der tonalen Klangsyntax*. Peter Lang Publishing.
- Elsisy, H., & Krishnan, A. (2008). Comparison of the acoustic and neural distortion product at 2f1-f2 in normal-hearing adults. *International Journal of Audiology*, 47(7), 431–438.
- Erlich, P. (1998). Tuning, tonality, and twenty-two-tone temperament. *Xenharmonikon*, 17, 12–40.
- Erlich, P. (2017).
- Fabian, D., Timmers, R., & Schubert, E. (2014). *Expressiveness in Music Performance: Empirical Approaches Across Styles and Cultures*. Oxford University Press, USA.
- Galbraith, G. C. (1994). Two-channel brain-stem frequency-following responses to pure tone and missing fundamental stimuli. *Electroencephalography and Clinical Neurophysiology/Evoked Potentials Section*, 92(4), 321–330.
- Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge university press.
- Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman; Hall/CRC.
- Giedraitis, K. (n.d.). Notation guide for edos 5-72 [Accessed: 2020-03-26]. [http://tallkite.com/misc\\_files/notation%5C%20guide%5C%20for%5C%20edos%5C%205-72.pdf](http://tallkite.com/misc_files/notation%5C%20guide%5C%20for%5C%20edos%5C%205-72.pdf)
- Greenberg, S., Marsh, J. T., Brown, W. S., & Smith, J. C. (1987). Neural temporal coding of low pitch. i. human frequency-following responses to complex tones. *Hearing research*, 25(2-3), 91–114.
- Guilford, J. P. (1954). System in the relationship of affective value to frequency and intensity of auditory stimuli. *The American Journal of Psychology*, 67(4), 691–695.
- Hába, A. (1927). *Neue Harmonielehre: Des Diatonischen-, Chromatischen-, Viertel-, Drittel-, Sechstel-, und Zwölftel-Tonsystems* (Vol. 3). Universal Edition.
- Hagle, T. M., & Mitchell, G. E. (1992). Goodness-of-fit measures for probit and logit. *American Journal of Political Science*, 36(3), 762–784.
- Hearne, G. M. (n.d.). Gareth Hearne [Accessed: 2020-03-25]. <https://soundcloud.com/gareth-hearne>

- Hearne, G. M. (2020a). Extended-diatonic interval names [Accessed: 2020-03-24]. [https://en.xen.wiki/w/Extended-diatonic\\_interval\\_names](https://en.xen.wiki/w/Extended-diatonic_interval_names)
- Hearne, G. M. (2020b). Shefkhed interval names [Accessed: 2020-03-24]. [https://en.xen.wiki/w/SHEFKHED\\_interval\\_names](https://en.xen.wiki/w/SHEFKHED_interval_names)
- Hearne, G. M., Milne, A. J., & Dean, R. T. (2019). Distributional analysis of n-dimensional feature space for 7-note scales in 22-tet, In *International Conference on Mathematics and Computation in Music*. Springer.
- Hoeschele, M., Weisman, R. G., & Sturdy, C. B. (2012). Pitch chroma discrimination, generalization, and transfer tests of octave equivalence in humans. *Attention, Perception, & Psychophysics*, 74(8), 1742–1760.
- Huron, D. (1991). Tonal consonance versus tonal fusion in polyphonic sonorities. *Music Perception: An Interdisciplinary Journal*, 9(2), 135–154.
- Huron, D. (1994). Interval-class content in equally tempered pitch-class sets: Common scales exhibit optimum tonal consonance. *Music Perception: An Interdisciplinary Journal*, 11(3), 289–305.
- Hutchinson, W., & Knopoff, L. (1978). The acoustic component of western consonance. *Journal of New Music Research*, 7(1), 1–29.
- Hyer, B. (2002). Tonality. In T. Christensen (Ed.), *The cambridge history of western music theory* (pp. 726–52). Cambridge, England, Cambridge University Press.
- Izumi, A. (2000). Japanese monkeys perceive sensory consonance of chords. *The Journal of the Acoustical Society of America*, 108(6), 3073–3078.
- Jeffreys, H. (1998). *The Theory of Probability* (Third) [originally published in 1961]. Oxford University Press.
- Johnson-Laird, P. N., Kang, O. E., & Leong, Y. C. (2012). On musical dissonance. *Music Perception: An Interdisciplinary Journal*, 30(1), 19–35.
- Johnston, B. (1984). *String Quartet no. 6*. Smith Publications.
- Kameoka, A., & Kuriyagawa, M. (1969a). Consonance theory part i: Consonance of dyads. *The Journal of the Acoustical Society of America*, 45(6), 1451–1459.
- Kameoka, A., & Kuriyagawa, M. (1969b). Consonance theory part ii: Consonance of complex tones and its calculation method. *The Journal of the Acoustical Society of America*, 45(6), 1460–1469.
- Kessler, E. J., Hansen, C., & Shepard, R. N. (1984). Tonal schemata in the perception of music in bali and in the west. *Music Perception: An Interdisciplinary Journal*, 2(2), 131–165.

- Knopoff, L., & Hutchinson, W. (1983). Entropy as a measure of style: The influence of sample length. *Journal of Music Theory*, 27(1), 75–97.
- Krishnan, A. (1999). Human frequency-following responses to two-tone approximations of steady-state vowels. *Audiology and Neurotology*, 4(2), 95–103.
- Krumhansl, C. L. (1990). Tonal hierarchies and rare intervals in music cognition. *Music Perception: An Interdisciplinary Journal*, 7(3), 309–324.
- Krumhansl, C. L. (2001). *Cognitive Foundations of Musical Pitch* (Vol. 17). Oxford University Press.
- Krumhansl, C. L., & Kessler, E. J. (1982). Tracing the dynamic changes in perceived tonal organization in a spatial representation of musical keys. *Psychological Review*, 89(4), 334.
- Krumhansl, C. L., Sandell, G. J., & Sergeant, D. C. (1987). The perception of tone hierarchies and mirror forms in twelve-tone serial music. *Music Perception: An Interdisciplinary Journal*, 5(1), 31–77.
- Krumhansl, C. L., & Schmuckler, M. A. (1986). The petroushka chord: A perceptual investigation. *Music Perception: An Interdisciplinary Journal*, 4(2), 153–184.
- Krumhansl, C. L., & Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *Journal of Experimental Psychology: Human Perception and Performance*, 5(4), 579.
- Kruschke, J. K., & Liddell, T. M. (2018). Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1), 155–177.
- Lantz, M. E., Kim, J.-K., & Cuddy, L. L. (2014). Perception of a tonal hierarchy derived from korean music. *Psychology of Music*, 42(4), 580–598.
- Lantz, M. E. (2002). *The Role of Duration and Frequency of Occurrence in Perceived Pitch Structure*. (Doctoral dissertation). ProQuest Information & Learning.
- Large, E. W. (2010). A dynamical systems approach to musical tonality. In *Nonlinear Dynamics in Human Behavior* (pp. 193–211). Springer.
- Large, E. W., Kim, J. C., Flaig, N. K., Bharucha, J. J., & Krumhansl, C. L. (2016). A neurodynamic account of musical tonality. *Music Perception: An Interdisciplinary Journal*, 33(3), 319–331.
- Lee, K. M., Skoe, E., Kraus, N., & Ashley, R. (2009). Selective subcortical enhancement of musical intervals in musicians. *Journal of Neuroscience*, 29(18), 5832–5840.
- Leman, M. (2000). An auditory model of the role of short-term memory in probe-tone ratings. *Music Perception: An Interdisciplinary Journal*, 17(4), 481–509.

- Lerdahl, F. (1988). Tonal pitch space. *Music Perception: An Interdisciplinary Journal*, 5(3), 315–349.
- Lerdahl, F. (1996). Calculating tonal tension. *Music Perception: An Interdisciplinary Journal*, 13(3), 319–363.
- Lerdahl, F., & Krumhansl, C. L. (2007). Modeling tonal tension. *Music Perception: An Interdisciplinary Journal*, 24(4), 329–366.
- Lerud, K. D., Almonte, F. V., Kim, J. C., & Large, E. W. (2014). Mode-locking neurodynamics predict human auditory brainstem responses to musical intervals. *Hearing Research*, 308, 41–49.
- Leung, Y., & Dean, R. T. (2018). Learning a well-formed microtonal scale: Pitch intervals and event frequencies. *Journal of New Music Research*, 47(3), 206–225.
- Loui, P. (2012). Learning and liking of melody and harmony: Further studies in artificial grammar learning. *Topics in Cognitive Science*, 4(4), 554–567.
- Loui, P., & Wessel, D. (2008). Learning and liking an artificial musical system: Effects of set size and repeated exposure. *Musicae Scientiae*, 12(2), 207–230.
- Loui, P., Wessel, D., & Kam, C. H. (2006). Acquiring new musical grammars: A statistical learning approach, In *28th Annual Conference of the Cognitive Science Society*.
- Loui, P., Wessel, D. L., & Kam, C. L. H. (2010). Humans rapidly learn grammatical structure in a new musical scale. *Music Perception: An Interdisciplinary Journal*, 27(5), 377–388.
- Madrid, A. L. (2015). *In Search of Julián Carrillo and Sonido 13*. Oxford University Press.
- Malmberg, C. F. (1918). The perception of consonance and dissonance. *Psychological Monographs*, 25(2), 93.
- Mandelbaum, J. (1961). *Multiple division of the octave and the tonal resources of the 19-tone equal temperament* (Doctoral dissertation). PhD Thesis. University of Indiana.
- Mashinter, K. (2006). Calculating sensory dissonance: Some discrepancies arising from the models of kameoka & kuriyagawa, and hutchinson & knopoff.
- Mathews, M. V., Pierce, J. R., Reeves, A., & Roberts, L. A. (1988). Theoretical and experimental explorations of the bohlen–pierce scale. *The Journal of the Acoustical Society of America*, 84(4), 1214–1222.

- McDaniel, C. N., & Williams, J. K. (2012). *The emancipation of consonance: A pedagogical approach to distinguishing between consonance and harmonic stability* (Doctoral dissertation). University of North Carolina at Greensboro.
- McDermott, J., Schultz, A., Undurraga, E., & Godoy, R. (2016). Consonance preferences are not universal: Indifference to dissonance among native amazonians. *The Journal of the Acoustical Society of America*, 139(4), 1994–1994.
- McKelvey, R. D., & Zavoina, W. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4(1), 103–120.
- Meyer, L. B. (2008). *Emotion and Meaning in Music* [originally published in 1956]. University of Chicago Press.
- Milne, A. J., Bulger, D., & Herff, S. A. (2017). Exploring the space of perfectly balanced rhythms and scales. *Journal of Mathematics and Music*, 11(2-3), 101–133.
- Milne, A. J., Bulger, D., Herff, S. A., & Sethares, W. A. (2015). Perfect balance: A novel principle for the construction of musical scales and meters, In *International Conference on Mathematics and Computation in Music*. Springer.
- Milne, A. J., & Dean, R. T. (2016). Computational creation and morphing of multi-level rhythms by control of evenness. *Computer Music Journal*, 40(1), 35–53.
- Milne, A. J., & Herff, S. A. (2020). The perceptual relevance of balance, evenness, and entropy in musical rhythms. *Cognition*, 203, 104233.
- Milne, A. J., & Holland, S. (2016). Empirically testing tonnetz, voice-leading, and spectral models of perceived triadic distance. *Journal of Mathematics and Music*, 10(1), 59–85.
- Milne, A. J., Laney, R., & Sharp, D. B. (2015). A spectral pitch class model of the probe tone data and scalic tonality. *Music Perception: An Interdisciplinary Journal*, 32(4), 364–393.
- Milne, A. J., Laney, R., & Sharp, D. B. (2016). Testing a spectral model of tonal affinity with microtonal melodies and inharmonic spectra. *Musicae Scientiae*, 20(4), 465–494.
- Milne, A. J., Sethares, W. A., Laney, R., & Sharp, D. B. (2011). Modelling the similarity of pitch collections with expectation tensors. *Journal of Mathematics and Music*, 5(1), 1–20.
- Mondor, T. A., & Morin, S. R. (2004). Primacy, recency, and suffix effects in auditory short-term memory for pure tones: Evidence from a probe recognition paradigm. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 58(3), 206.

- Monzo, J. (n.d.). Equal-temperament [Accessed: 2021-01-05]. <http://tonalsoft.com/enc/e/equal-temperament.aspx>
- Moore, B. C. (1973). Frequency difference limens for short-duration tones. *The Journal of the Acoustical Society of America*, 54(3), 610–619.
- Moore, B. C., Glasberg, B. R., & Shailer, M. J. (1984). Frequency and intensity difference limens for harmonics within complex tones. *The Journal of the Acoustical Society of America*, 75(2), 550–561.
- Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PloS One*, 9(2), e89642.
- Nam, U. (1998). Pitch distributions in korean court music: Evidence consistent with tonal hierarchies. *Music Perception: An Interdisciplinary Journal*, 16(2), 243–247.
- Op de Coul, M. (2020). Scala [computer software] [Version 2.44p]. <http://www.huygens-fokker.org/scala/>
- Oram, N., & Cuddy, L. L. (1995). Responsiveness of western adults to pitch-distributional information in melodic sequences. *Psychological Research*, 57(2), 103–118.
- Pandya, P. K., & Krishnan, A. (2004). Human frequency-following response correlates of the distortion product at 2f<sub>1</sub>-f<sub>2</sub>. *Journal of the American Academy of Audiology*, 15(3), 184–197.
- Parham, L. C. C. (1987). *Evaluative ratings and exploration of sounds of differing pitch and complexity* (Doctoral dissertation). University of Toronto.
- Parncutt, R. (2011). The tonic as triad: Key profiles as pitch salience profiles of tonic triads. *Music Perception: An Interdisciplinary Journal*, 28(4), 333–366.
- Parncutt, R., & Cohen, A. J. (1995). Identification of microtonal melodies: Effects of scale-step size, serial order, and training. *Perception & Psychophysics*, 57(6), 835–846.
- Parncutt, R., & Hair, G. (2011). Consonance and dissonance in music theory and psychology: Disentangling dissonant dichotomies. *Journal of Interdisciplinary Music Studies*, 5(2).
- Parncutt, R., Reisinger, D., Fuchs, A., & Kaiser, F. (2019). Consonance and prevalence of sonorities in western polyphony: Roughness, harmonicity, familiarity, evenness, diatonicity. *Journal of New Music Research*, 48(1), 1–20.
- Partch, H. (1949). *Genesis of a Music: Monophony: The Relation of its Music to Historic and Contemporary Trends; its Philosophy, Concepts, and Principles; its Relation*

- to Historic and Proposed Intonations; and its Application to Musical Instruments.* University of Wisconsin Press.
- Pearce, M. T. (2005). *The construction and evaluation of statistical models of melodic structure in music perception and composition* (Doctoral dissertation). City University London.
- Pickles, J. (1988). *An Introduction to the Physiology of Hearing.* Academic Press.
- Plack, C. J. (2010). Musical consonance: The importance of harmonicity. *Current Biology*, 20(11), R476–R478.
- Plack, C. J., Oxenham, A. J., & Fay, R. R. (2006). *Pitch: Neural Coding and Perception* (Vol. 24). Springer Science & Business Media.
- Plomp, R., & Levelt, W. J. M. (1965). Tonal consonance and critical bandwidth. *The Journal of the Acoustical Society of America*, 38(4), 548–560.
- Regnault, P., Bigand, E., & Besson, M. (2001). Different brain mechanisms mediate sensitivity to sensory consonance and harmonic context: Evidence from auditory event-related brain potentials. *Journal of Cognitive Neuroscience*, 13(2), 241–255.
- Rickman, M. D., Chertoff, M. E., & Hecox, K. E. (1991). Electrophysiological evidence of nonlinear distortion products to two-tone stimuli. *The Journal of the Acoustical Society of America*, 89(6), 2818–2826.
- Roberts, L. A. (1986). Consonance judgements of musical chords by musicians and untrained listeners. *Acta Acustica United with Acustica*, 62(2), 163–171.
- Roberts, L. A., & Shaw, M. L. (1984). Perceived structure of triads. *Music Perception: An Interdisciplinary Journal*, 2(1), 95–124.
- Rothenberg, D. (1975). A mathematical model for perception applied to the perception of pitch. In *Formal Aspects of Cognitive Processes* (pp. 126–141). Springer.
- Rothenberg, D. (1977). A model for pattern perception with musical applications part ii: The information content of pitch structures. *Mathematical Systems Theory*, 11(1), 353–372.
- Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65.
- Schellenberg, E. G., & Trainor, L. J. (1996). Sensory consonance and the perceptual similarity of complex-tone harmonic intervals: Tests of adult and infant listeners. *The Journal of the Acoustical Society of America*, 100(5), 3321–3328.



- Schmuckler, M. A. (1989). Expectation in music: Investigation of melodic and harmonic processes. *Music Perception: An Interdisciplinary Journal*, 7(2), 109–149.
- Sethares, W. A. (1993). Local consonance and the relationship between timbre and scale. *The Journal of the Acoustical Society of America*, 94(3), 1218–1228.
- Sethares, W. A. (2005). *Tuning, Timbre, Spectrum, Scale*. Springer Science & Business Media.
- Smit, E. A., Milne, A. J., Dean, R. T., & Weidemann, G. (2019). Perception of affect in unfamiliar musical chords. *PloS One*, 14(6), e0218570.
- Smith, N. A., & Cuddy, L. L. (2003). Perceptions of musical dimensions in beethoven's waldslein sonata: An application of tonal pitch space theory. *Musicae Scientiae*, 7(1), 7–34.
- Smith, N. A., & Schmuckler, M. A. (2004). The perception of tonal structure through the differentiation and organization of pitches. *Journal of Experimental Psychology: Human Perception and Performance*, 30(2), 268.
- Stanford, S., Milne, A., & MacRitchie, J. (2018). The effect of isomorphic pitch layouts on the transfer of musical learning. *Applied Sciences*, 8(12), 2514.
- Steinbeis, N., Koelsch, S., & Sloboda, J. A. (2006). The role of harmonic expectancy violations in musical emotions: Evidence from subjective, physiological, and neural responses. *Journal of Cognitive Neuroscience*, 18(8), 1380–1393.
- Stolzenburg, F. (2015). Harmony perception by periodicity detection. *Journal of Mathematics and Music*, 9(3), 215–238.
- Strasburger, H., & Parncutt, H. (1994). Applying psychoacoustics in composition: Harmonic progressions of non-harmonic sonorities. *Perspectives of New Music*, 32(1), 88–129.
- Temperley, D. (n.d.). Harmonic analyses [Accessed: 2018-06-06]. [http://rockcorpus.midside.com/harmonic\\_analyses.html](http://rockcorpus.midside.com/harmonic_analyses.html)
- Temperley, D., & de Clercq, T. (2013). Statistical analysis of harmony and melody in rock music. *Journal of New Music Research*, 42(3), 187–204.
- Terhardt, E. (1984). The concept of musical consonance: A link between music and psychoacoustics. *Music Perception: An Interdisciplinary Journal*, 1(3), 276–295.
- Tillmann, B., & Lebrun-Guillaud, G. (2006). Influence of tonal and temporal expectations on chord processing and on completion judgments of chord sequences. *Psychological Research*, 70(5), 345–358.

- Toiviainen, P., & Krumhansl, C. L. (2003). Measuring and modeling real-time responses to music: The dynamics of tonality induction. *Perception*, 32(6), 741–766.
- Trainor, L. J., Tsang, C. D., & Cheung, V. H. (2002). Preference for sensory consonance in 2- and 4-month-old infants. *Music Perception: An Interdisciplinary Journal*, 20(2), 187–194.
- van de Geer, J. P., Levelt, W. J., & Plomp, R. (1962). The connotation of musical consonance.
- Veall, M. R., & Zimmermann, K. F. (1992). Pseudo-R<sup>2</sup>'s in the ordinal probit model. *Journal of Mathematical Sociology*, 16(4), 333–342.
- Veall, M. R., & Zimmermann, K. F. (1994). Evaluating pseudo-R<sup>2</sup>'s for binary probit models. *Quality and Quantity*, 28(2), 151–164.
- Vega, D. (2003). A perceptual experiment on harmonic tension and melodic attraction in Ierdlahl's tonal pitch space. *Musicae Scientiae*, 7(1), 35–55.
- Vehtari, A. (2020). Cross-validation faq [Accessed: 2020-12-17]. <https://avehtari.github.io/modelselection/CV-FAQ.html>
- Vicentino, N. (1996). *Ancient Music Adapted to Modern Practice*. Yale University Press.
- Vos, J. (1982). The perception of pure and mistuned musical fifths and major thirds: Thresholds for discrimination, beats, and identification. *Perception & Psychophysics*, 32(4), 297–313.
- Wallin, N. L., Merker, B., & Brown, S. (2001). *The Origins of Music*. MIT press.
- West, R. J., & Fryer, R. (1990). Ratings of suitability of probe tones as tonics after random orderings of notes of the diatonic scale. *Music Perception: An Interdisciplinary Journal*, 7(3), 253–258.
- Wilson, E. (1975a). *Letter to John Chalmers pertaining to Moments of Symmetry/Tanabe Cycle*. [Accessed: 2019-01-10]. <http://www.anphoria.com/meruone.pdf>
- Wilson, E. (1975b). On the development of intonational systems by extended linear mapping. *Xenharmonikon*, 3.
- Würschmidt, J. (1921). Die 19-stufige skala. eine nat'urliche erweiterung unseres tonsystems. *Neue Musik-Zeitung*, 42, 215–16.
- Youngblood, J. E. (1958). Style as information. *Journal of Music Theory*, 2(1), 24–35.
- Zatorre, R. J., Delhommeau, K., & Zarate, J. M. (2012). Modulation of auditory cortex response to pitch variation following training with microtonal melodies. *Frontiers in Psychology*, 3, 544.

# Appendix

## Appendix A

# Formal Specification of the Spectral Pitch Class Similarity Model

In this section, adapted from the Appendix of Milne, Laney, et al. (2015), we give a formal mathematical specification of our model. The techniques used are based on those introduced by Milne et al. (2011). The MATLAB routines that embody these routines can be downloaded from [http://www.dynamictonality.com/probe\\_tone\\_files/](http://www.dynamictonality.com/probe_tone_files/).

Let a chord comprising  $M$  tones, each of which contains  $N$  partials, be represented by the matrix  $\mathbf{X}_f \in \mathbb{R}^{M \times N}$ . Each row of  $\mathbf{X}_f$  represents a tone in the chord, and each element of the row is the frequency of a partial of that tone. In our model, we use the first 35 partials (so  $N = 35$ ); this means that, if  $\mathbf{X}_f$  is a three-tone chord, it will be a  $3 \times 35$  matrix.

The first step is to convert the partials' frequencies into pitch class cents values:

$$x_{\text{pc}}[m, n] = 1200 \lceil \log_2(x_f[m, n] / x_{\text{ref}}) \rceil \bmod 1200, \quad (\text{A.1})$$

where  $\lceil \cdot \rceil$  is the nearest integer function, and  $x_{\text{ref}}$  is an arbitrary reference frequency (e.g., the frequency of middle C). These values are then collected into a single *pitch class vector* denoted  $\tilde{\mathbf{x}}_{\text{pc}} \in \mathbb{Z}^{35M}$  indexed by  $i$  such that  $x_{\text{pc}}[m, n] \mapsto \tilde{x}_{\text{pc}}[i]$ , where  $i = (m - 1)N + n$ .

Let each of the partials have an associated weight  $x_w[m, n]$ , which represents their *salience*, or probability of being perceived. The saliences of the tonic triad are parameterized by a *roll-off* value  $\rho \in \mathbb{R}$ , so that

$$x_w[m, n] = n^{-\rho} \quad m = 1, \dots, M, \text{ and } n = 1, \dots, 35, \quad (\text{A.2})$$

When  $\rho = 0$ , all partials of a tone  $m$  have a weight of 1; as  $\rho$  increases, the weights of its higher partials are reduced (see the final paragraph of this appendix for our choice of  $\rho$  and the reasoning behind it). These values are collected into a single *weighting vector*  $\tilde{\mathbf{x}}_w \in \mathbb{R}^{35M}$  also indexed by  $i$  such that  $x_w[m, n] \mapsto \tilde{x}_w[i]$ , where  $i = (m - 1)N + n$  (the precise method used to reshape the matrix into vector form is unimportant so long as it matches that used for the pitch class vector).

The partials (their pitch classes and weights in  $\tilde{\mathbf{x}}_{pc}$  and  $\tilde{\mathbf{x}}_w$ ) are embedded in a *spectral pitch class salience matrix*  $\mathbf{X}_{pcs} \in \mathbb{R}^{35N \times 1200}$  indexed by  $i$  and  $j$ :

$$x_{pcs}[i, j] = \tilde{x}_w[i] \delta[j - \tilde{x}_{pc}[i]] \quad i = 1, \dots, 35N, \text{ and } j = 0, \dots, 1199, \quad (\text{A.3})$$

where  $\delta[z]$  is the Kronecker delta function, which equals 1 when  $z = 0$ , and equals 0 when  $z \neq 0$ . This equation means that the matrix  $\mathbf{X}_{pcs}$  is all zeros except for  $35N$  elements, and each element indicates the salience  $x_{pcs}[i, j]$  of partial  $i$  at pitch  $j$ .

To model the uncertainty of pitch perception, these  $35N$  delta ‘spikes’ are ‘smeared’ by circular convolution with a discrete Gaussian kernel  $\mathbf{g}$ , which is also indexed by  $j$ , and is parameterized with a *smoothing* standard deviation  $\sigma \in [0, \infty)$  to give a *spectral pitch class response matrix*  $\mathbf{X}_{pcr} \in \mathbb{R}^{35N \times 1200}$ , which is indexed by  $i$  and  $k$ :

$$\mathbf{x}_{pcr}[i] = \mathbf{x}_{pcs}[i] * \mathbf{g}, \quad (\text{A.4})$$

where  $\mathbf{x}_{pcr}[i]$  is the  $i$ th row of  $\mathbf{X}_{pcr}$ , and  $*$  denotes circular convolution over the period of 1200 cents; that is,

$$x_{pcr}[i, k] = \sum_{j=0}^{1199} x_{pcs}[i, j] g[(k - j) \bmod 1200] \\ i = 1, \dots, 35N, \text{ and } k = 0, \dots, 1199. \quad (\text{A.5})$$

In our implementation, we make use of the circular convolution theorem, which allows (A.4) to be calculated efficiently with fast Fourier transforms; that is,  $\mathbf{f} * \mathbf{g} = \mathcal{F}^{-1}(\mathcal{F}(\mathbf{f}) \circ \mathcal{F}(\mathbf{g}))$ , where  $*$  is circular convolution,  $\mathcal{F}$  denotes the Fourier transform,  $\circ$  is the Hadamard (elementwise) product, and  $\mathbf{f}$  stands for  $\mathbf{x}_{pcs}[i]$ .

Equation (A.4) can be interpreted as adding random noise (with a Gaussian distribution) to the original pitch classes in  $\mathbf{X}_{pcs}$ , thereby simulating perceptual pitch uncertainty. The standard deviation of the Gaussian distribution  $\sigma$  models the pitch

difference limen (just noticeable difference) (Milne et al., 2011, Online Supplementary, App. A). In laboratory experiments with sine waves, the pitch difference limen is approximately 3 cents in the central range of frequency (Moore, 1973; Moore et al., 1984). We would expect the pitch difference limen in the more distracting setting of listening to music to be somewhat wider. Indeed, the value of  $\sigma$  was optimized—with respect to the probe tone data—at approximately 6 cents.

Each element  $x_{\text{pcr}}[i, k]$  of this matrix models the probability of the  $i$ th partial in  $x_{\text{pc}}$  being perceived at pitch class  $k$ . In order to summarize the responses to all the pitches, we take the column sum, which gives a vector of the expected numbers of partials perceived at pitch class  $k$ . This 1200-element row vector is denoted a *spectral pitch class vector*  $\mathbf{x}$ :

$$\mathbf{x} = \mathbf{1}'\mathbf{X}_{\text{pcr}}, \quad (\text{A.6})$$

where  $\mathbf{1}'$  denotes a row vector of  $35N$  ones. The spectral pitch class similarity of two such vectors  $\mathbf{x}$  and  $\mathbf{y}$  is given by any standard similarity metric. We choose the cosine:

$$s(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{xy}'}{\sqrt{(\mathbf{xx}')(\mathbf{yy}')}}, \quad (\text{A.7})$$

where  $'$  denotes the matrix transpose operator that turns a row vector into a column vector (and vice versa). Because  $\mathbf{x}$  and  $\mathbf{y}$  contain only non-negative values, their cosine similarity falls between 0 and 1, where 1 implies the two vectors are parallel, and 0 implies they are orthogonal.

We use this model to establish the similarities of a variety of probes with respect to a context. Let the context be represented by the spectral pitch class vector  $\mathbf{x}$ , and let the  $P$  different probes  $\mathbf{y}_p$  be collected into a matrix of spectral pitch class vectors denoted  $\mathbf{Y} \in \mathbb{R}^{P \times 1200}$ . The column vector of  $P$  similarities between each of the probes and the context is then denoted  $\mathbf{s}(\mathbf{x}, \mathbf{Y}) \in \mathbb{R}^P$ . For example, the context may be a major triad built from harmonic complex tones and the probes may be single harmonic complex tones at the twelve chromatic pitches. In this case, the 105 harmonics from the context (35 partials for each of the three different chord tones) are embedded into a single spectral pitch class vector  $\mathbf{x}$ , as described in (A.1–A.6).

Each of the twelve (for Experiments 1-2) or twenty-two (for Experiments 3-5) differently pitched probe tones' 35 harmonics are embedded into twelve or twenty-two spectral pitch class vectors  $\mathbf{y}_p$  respectively. The similarities of the context and

the probes are calculated—as described in (A.7)—to give the vector of their similarities  $\mathbf{s}(\mathbf{x}, \mathbf{Y})$ .

In analysis of all experiments in this thesis, the  $\rho$  value of  $5/3$  is used. This is the value of the roll-off of the partials used in the synthesis of the experimental stimulus. In Milne, Laney, et al. (2015) the value of  $\sigma$  was optimized, iteratively, to minimize the sum of squared residuals between the model's predictions and the empirical data. The same value of  $\sigma$ , 10 cents, was found to result in a better performing model for the analysis of both the probe tones and triads of Experiment 1 of this thesis than values of 14 cents and 6 cents, and a  $\sigma$  value of 10 cents was then used in the analysis of the remaining experiments.

## Appendix B

# Experiments 1 & 2 Tones

## B.1 Experiment 1

### B.1.1 Pre-registered model

Significant as main effects in this model as well as SPCS (medium effect size, evid. ratio > 11999) and Prevalence (medium, evid. ratio > 11999) in the positive direction, are Recency (large, evid. ratio 799) and RelHeight (medium, evid. ratio > 11999). Interactions of MusSoph with Recency (medium, evid. ratio 332.33), SPCS (small, evid. ratio 856.14), Prevalence (small, evid. ratio 5999) and MelCont (very small, evid. ratio 76.92) in the positive direction are also significant, confirming our third hypothesis.

The same model was run without SPCS and Prevalence as predictors, and, judged by LOOIC, then performed significantly worse (difference in LOOIC of 177.89 with a standard error of 21.64). Our hypothesis is thus confirmed: Perceived goodness-of-fit and perceived stability of probe tones may be modelled by the SPCS between the pitches of the context and the probe and the statistical prevalence in Western music of the probe within the context scale.

This model, with a Pseudo- $R^2$  value of 0.51, does not fit as well as Model 2.1. The *ppcheck* plot for this model resembles that of Model 2.1.



TABLE B.1: Pre-registered probe tones model significant population-level Effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.89	0.13	-3.16	-2.64	6590	1.00
Intercept[2]	-1.28	0.12	-1.52	-1.04	6431	1.00
Intercept[3]	-0.04	0.12	-0.28	0.19	6329	1.00
Intercept[4]	0.68	0.12	0.45	0.92	6394	1.00
Intercept[5]	1.91	0.12	1.66	2.15	6704	1.00
Intercept[6]	3.46	0.14	3.19	3.73	7225	1.00
MusSoph	-0.12	0.10	-0.32	0.07	5520	1.00
RelHeight	0.36	0.07	0.23	0.49	8868	1.00
Recency	0.78	0.25	0.29	1.27	11469	1.00
SPCS	0.51	0.09	0.33	0.70	10323	1.00
MelCont	-0.06	0.05	-0.16	0.03	16138	1.00
Prevalence	0.50	0.09	0.33	0.66	10653	1.00
MusSoph:Recency	0.58	0.21	0.17	1.00	10953	1.00
MusSoph:SPCS	0.24	0.08	0.09	0.40	10196	1.00
MusSoph:MelCont	0.08	0.04	0.01	0.15	16922	1.00
MusSoph:Prevalence	0.25	0.07	0.11	0.40	11371	1.00

Significant population-level effects for the preregistered probe tones model along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

## B.1.2 Model 2.1

TABLE B.2: All population-level effects for Model 2.1, which was summarized in Table 2.2

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.19	0.14	-3.48	-2.91	6404	1
Intercept[2]	-1.45	0.13	-1.71	-1.19	6603	1
Intercept[3]	-0.11	0.13	-0.37	0.14	6655	1
Intercept[4]	0.67	0.13	0.42	0.92	6666	1
Intercept[5]	2.01	0.13	1.74	2.27	6795	1
Intercept[6]	3.68	0.15	3.39	3.97	7386	1
MusSoph	-0.13	0.10	-0.34	0.07	5142	1
Height	0.18	0.08	0.02	0.35	6875	1
RelHeight	0.27	0.07	0.13	0.42	9443	1
Height <sup>2</sup>	0.01	0.06	-0.12	0.13	8135	1
RelHeight <sup>2</sup>	-0.12	0.06	-0.23	-0.01	10676	1
Primacy	0.14	0.18	-0.21	0.50	13637	1
Previous	0.35	0.07	0.22	0.48	8957	1
Recency	0.73	0.27	0.20	1.28	7871	1
SPCS	0.58	0.10	0.39	0.77	7803	1
Prevalence	0.52	0.09	0.34	0.70	8553	1
MelCont	-0.02	0.05	-0.12	0.07	12015	1
Count	-0.07	0.11	-0.28	0.14	9803	1
TrialNo	-0.01	0.07	-0.15	0.12	9806	1
InContTrialNo	0.03	0.06	-0.08	0.14	12898	1
TrialNo <sup>2</sup>	0.04	0.06	-0.08	0.17	10005	1
BlockOrder	0.05	0.23	-0.4	0.51	4907	1
MusSoph:Height	0.01	0.07	-0.12	0.15	7185	1
MusSoph:RelHeight	-0.09	0.06	-0.20	0.02	10058	1
MusSoph:Height <sup>2</sup>	0.02	0.05	-0.08	0.12	7742	1
MusSoph:RelHeight <sup>2</sup>	-0.01	0.04	-0.09	0.08	14314	1
MusSoph:Primacy	-0.16	0.13	-0.43	0.10	13153	1
MusSoph:Previous	-0.11	0.05	-0.22	0.00	7853	1
MusSoph:Recency	0.66	0.23	0.22	1.14	8186	1
MusSoph:SPCS	0.26	0.09	0.09	0.42	8752	1
MusSoph:Prevalence	0.28	0.08	0.13	0.43	8281	1
MusSoph:MelCont	0.07	0.04	0.00	0.15	14239	1
MusSoph:Count	0.02	0.06	-0.09	0.13	11649	1
Height:TrialNo	0.05	0.05	-0.04	0.14	12827	1
RelHeight:TrialNo	-0.01	0.05	-0.1	0.08	12354	1
Height <sup>2</sup> :TrialNo	0.07	0.04	0.00	0.14	12299	1
RelHeight <sup>2</sup> :TrialNo	0.01	0.04	-0.07	0.09	15186	1
Primacy:TrialNo	0.03	0.13	-0.23	0.29	16351	1
Previous:TrialNo	0.03	0.04	-0.04	0.11	13771	1
Recency:TrialNo	0.24	0.14	-0.03	0.51	15574	1
SPCS:TrialNo	-0.1	0.05	-0.2	0.00	10157	1
Prevalence:TrialNo	0.01	0.05	-0.09	0.11	11262	1
MelCont:TrialNo	-0.03	0.05	-0.12	0.06	11292	1
Count:TrialNo	0.04	0.05	-0.05	0.13	10880	1
Height:InContTrialNo	-0.05	0.05	-0.14	0.04	12907	1
RelHeight:InContTrialNo	0.08	0.04	-0.01	0.16	14115	1
Height <sup>2</sup> :InContTrialNo	-0.06	0.04	-0.13	0.01	12807	1
RelHeight <sup>2</sup> :InContTrialNo	0.00	0.04	-0.08	0.08	12696	1

*Continued on next page*

Table B.2 – Continued from previous page

Effect	Estimate	Est. Error	l-95% CI	u-95% CI	Eff. Sample	Rhat
Primacy:InContTrialNo	0.09	0.14	-0.18	0.37	14049	1
Previous:InContTrialNo	0.04	0.04	-0.04	0.12	13963	1
Recency:InContTrialNo	0.15	0.14	-0.12	0.42	16684	1
SPCS:InContTrialNo	0.06	0.05	-0.03	0.15	11313	1
Prevalence:InContTrialNo	0.03	0.05	-0.06	0.13	12750	1
MelCont:InContTrialNo	0.01	0.04	-0.06	0.09	14710	1
Count:InContTrialNo	-0.04	0.04	-0.13	0.04	12492	1
Height: TrialNo <sup>2</sup>	-0.07	0.05	-0.17	0.03	10859	1
RelHeight: TrialNo <sup>2</sup>	-0.01	0.05	-0.1	0.09	9761	1
Height <sup>2</sup> : TrialNo <sup>2</sup>	-0.06	0.04	-0.13	0.02	11354	1
RelHeight <sup>2</sup> : TrialNo <sup>2</sup>	0.00	0.04	-0.08	0.09	10455	1
Primacy: TrialNo <sup>2</sup>	-0.02	0.14	-0.3	0.26	12349	1
Previous: TrialNo <sup>2</sup>	-0.04	0.04	-0.11	0.04	12943	1
Recency: TrialNo <sup>2</sup>	0.08	0.15	-0.22	0.38	13194	1
SPCS: TrialNo <sup>2</sup>	0.04	0.06	-0.08	0.15	10101	1
Prevalence: TrialNo <sup>2</sup>	-0.04	0.05	-0.15	0.06	11872	1
MelCont: TrialNo <sup>2</sup>	0.04	0.04	-0.04	0.12	11624	1
Count: TrialNo <sup>2</sup>	-0.01	0.05	-0.11	0.08	10802	1
Height: BlockOrder	-0.01	0.14	-0.29	0.27	6411	1
RelHeight: BlockOrder	0.04	0.11	-0.18	0.25	11668	1
Height <sup>2</sup> : BlockOrder	0.03	0.11	-0.18	0.24	7407	1
RelHeight <sup>2</sup> : BlockOrder	0.10	0.08	-0.06	0.26	13795	1
Primacy: BlockOrder	-0.11	0.26	-0.62	0.41	14552	1
Previous: BlockOrder	-0.1	0.11	-0.32	0.12	8114	1
Recency: BlockOrder	0.24	0.46	-0.66	1.14	8175	1
SPCS: BlockOrder	-0.25	0.17	-0.59	0.08	8055	1
Prevalence: BlockOrder	-0.12	0.15	-0.43	0.18	7582	1
MelCont: BlockOrder	0.09	0.08	-0.06	0.24	12333	1
Count: BlockOrder	0.19	0.20	-0.19	0.57	10966	1

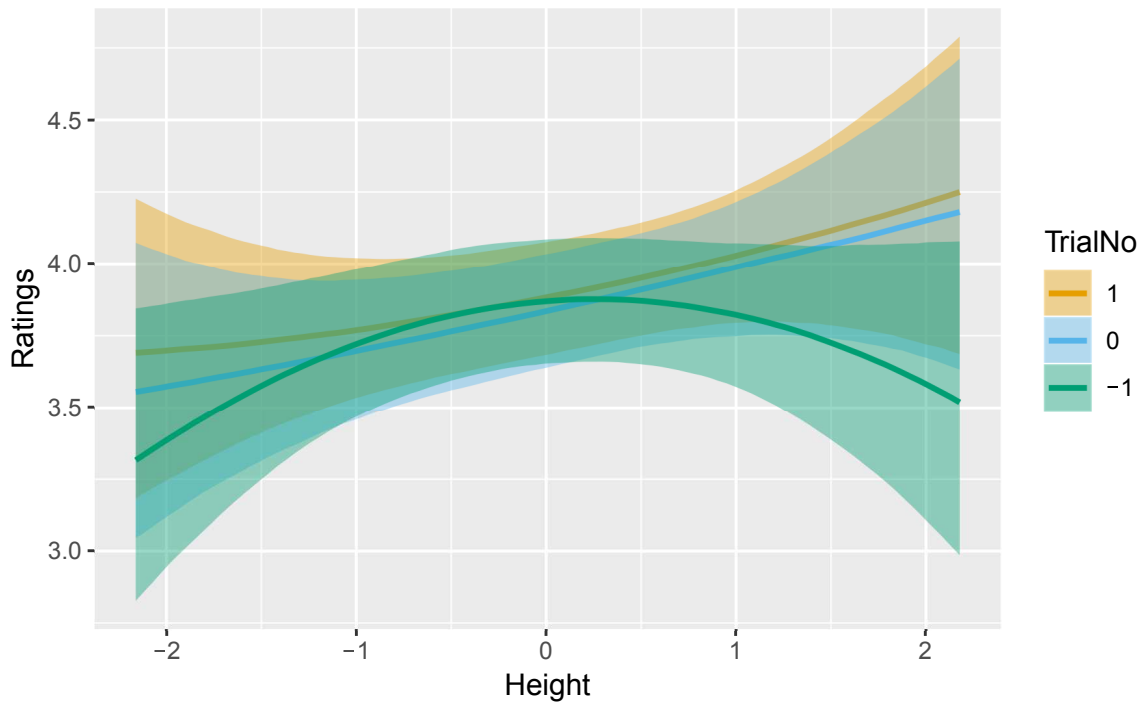


FIGURE B.1: Conditional effect of Height: TrialNo for Model 2.1

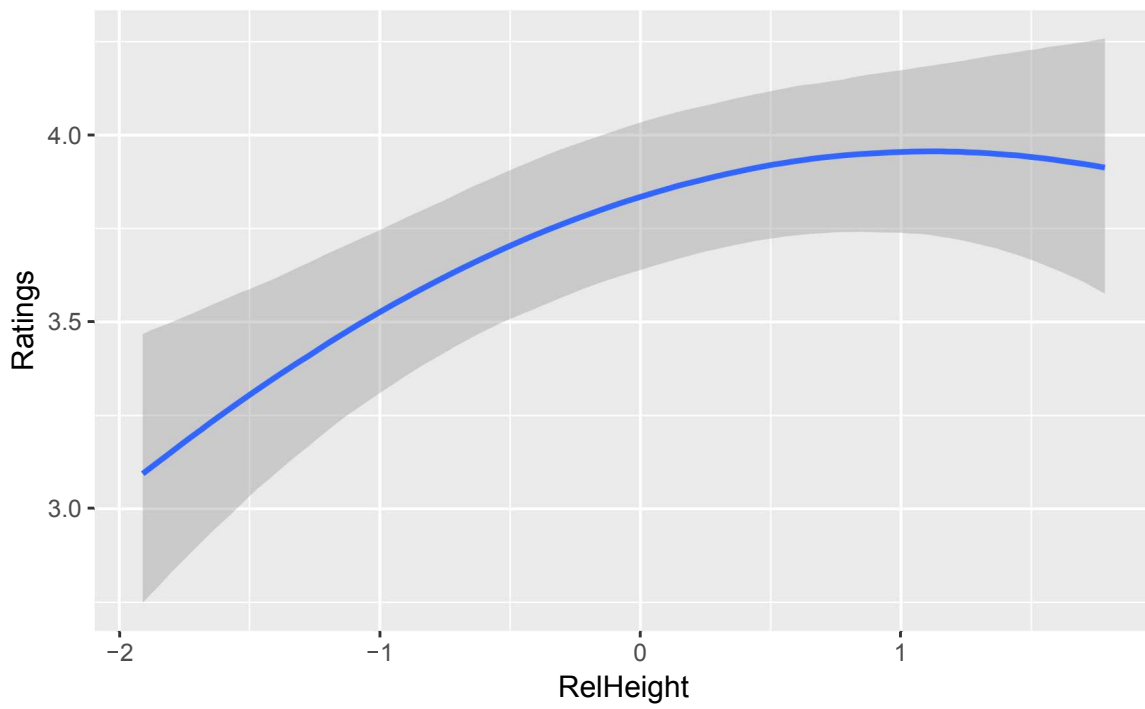


FIGURE B.2: Conditional effect of RelHeight for Model 2.2

TABLE B.3: Model 2.1 but without Prevalence significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.08	0.14	-3.35	-2.82	1308	1
Intercept[2]	-1.38	0.13	-1.64	-1.13	1373	1
Intercept[3]	-0.09	0.13	-0.33	0.16	1372	1
Intercept[4]	0.66	0.13	0.42	0.91	1350	1
Intercept[5]	1.92	0.13	1.65	2.17	1486	1
Intercept[6]	3.46	0.14	3.19	3.74	1544	1
MusSoph	-0.12	0.10	-0.32	0.08	682	1
Height	0.15	0.09	-0.02	0.32	1364	1
RelHeight	0.26	0.07	0.12	0.40	1682	1
Height <sup>2</sup>	0.00	0.06	-0.12	0.11	1194	1
RelHeight <sup>2</sup>	-0.11	0.06	-0.22	-0.01	1896	1
Previous	0.32	0.06	0.19	0.45	1717	1
Recency	0.61	0.26	0.12	1.14	1306	1
SPCS	0.88	0.09	0.70	1.07	1425	1
TrialNo	-0.01	0.06	-0.14	0.11	1715	1
lnContTrialNo	0.04	0.06	-0.07	0.15	2517	1
MusSoph:Previous	-0.12	0.05	-0.23	-0.02	1439	1
MusSoph:Recency	0.59	0.23	0.15	1.04	1381	1
MusSoph:SPCS	0.41	0.09	0.23	0.59	1190	1
Height:TrialNo	0.04	0.05	-0.05	0.13	2474	1
Height <sup>2</sup> :TrialNo	0.07	0.03	0.00	0.14	3063	1
SPCS:TrialNo	-0.08	0.04	-0.16	-0.01	2208	1
SPCS:lnContTrialNo	0.08	0.04	0.00	0.16	2163	1

Significant population-level effects for Model 2.1 but without Prevalence along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

TABLE B.4: Model 2.1 but with ScaleTone instead of SPCS significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.53	0.17	-2.86	-2.19	975	1.00
Intercept[2]	-0.81	0.16	-1.12	-0.50	961	1.00
Intercept[3]	0.52	0.16	0.21	0.84	968	1.00
Intercept[4]	1.31	0.16	1.00	1.63	971	1.00
Intercept[5]	2.65	0.16	2.34	2.98	991	1.00
Intercept[6]	4.34	0.18	4.00	4.68	1040	1.00
MusSoph	-0.40	0.13	-0.66	-0.16	745	1.00
Height	0.17	0.08	0.01	0.33	909	1.00
RelHeight	0.27	0.07	0.12	0.41	1281	1.00
RelHeight <sup>2</sup>	-0.12	0.06	-0.23	-0.02	1433	1.00
Previous	0.33	0.06	0.20	0.45	1363	1.00
Recency	0.74	0.27	0.22	1.28	906	1.00
Prevalence	0.60	0.09	0.43	0.78	901	1.00
ScaleTone	1.10	0.18	0.75	1.46	933	1.00
MusSoph:Previous	-0.11	0.05	-0.22	-0.01	1154	1.01
MusSoph:Recency	0.67	0.24	0.20	1.14	1202	1.00
MusSoph:Prevalence	0.33	0.08	0.17	0.49	958	1.00
MusSoph:MelCont	0.07	0.04	0.00	0.15	2191	1.00
MusSoph:ScaleTone	0.44	0.15	0.14	0.75	828	1.00

Significant population-level effects for Model 2.1 but with ScaleTone instead of SPCS along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ )'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

TABLE B.5: Model 2.1 but without Prevalence and with ScaleTone instead of SPCS significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.05	0.17	-2.38	-1.71	777	1.00
Intercept[2]	-0.41	0.17	-0.73	-0.07	774	1.00
Intercept[3]	0.86	0.17	0.55	1.20	778	1.00
Intercept[4]	1.60	0.17	1.28	1.96	795	1.00
Intercept[5]	2.84	0.17	2.50	3.20	797	1.00
Intercept[6]	4.36	0.18	4.01	4.72	849	1.00
MusSoph	-0.57	0.14	-0.84	-0.30	526	1.01
Height	0.14	0.08	-0.01	0.30	1124	1.00
RelHeight	0.26	0.07	0.12	0.38	1423	1.00
Height <sup>2</sup>	0.01	0.06	-0.12	0.13	1117	1.00
RelHeight <sup>2</sup>	-0.12	0.06	-0.23	-0.01	1962	1.00
Previous	0.30	0.06	0.18	0.43	1529	1.00
Recency	0.62	0.26	0.12	1.14	1289	1.00
ScaleTone	1.64	0.18	1.30	2.01	964	1.00
lnContTrialNo	-0.06	0.06	-0.19	0.06	2233	1.00
BlockOrder	0.36	0.28	-0.18	0.92	798	1.00
MusSoph:Height	0.01	0.07	-0.12	0.14	1150	1.00
MusSoph:RelHeight	-0.09	0.05	-0.19	0.01	1772	1.00
MusSoph:Height <sup>2</sup>	0.02	0.05	-0.08	0.12	959	1.00
MusSoph:RelHeight <sup>2</sup>	-0.01	0.04	-0.09	0.07	2419	1.00
MusSoph:Primacy	-0.18	0.13	-0.45	0.07	2178	1.00
MusSoph:Previous	-0.13	0.05	-0.24	-0.03	1257	1.00
MusSoph:Recency	0.62	0.24	0.15	1.08	1361	1.00
MusSoph:ScaleTone	0.75	0.16	0.44	1.07	839	1.00
Height:lnContTrialNo	-0.06	0.05	-0.15	0.03	1895	1.00
Height <sup>2</sup> :lnContTrialNo	-0.07	0.03	-0.13	-0.00	2411	1.00
ScaleTone:BlockOrder	-0.67	0.32	-1.26	-0.06	707	1.01

Significant population-level effects for Model 2.1 but without Prevalence and with ScaleTone instead of SPCS along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ ). The chains are obtained via *Markov chain Monte Carlo* (MCMC), which are a class of algorithms used in *brms* for sampling from a probability distribution.

## B.2 Experiment 2

### B.2.1 Model 2.2

TABLE B.6: All population-level effects for Model 2.2, which was summarized in Table 2.5

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.27	0.11	-3.49	-3.05	5216	1
Intercept[2]	-2.09	0.10	-2.29	-1.88	5043	1
Intercept[3]	-0.93	0.10	-1.13	-0.73	5014	1
Intercept[4]	-0.16	0.10	-0.36	0.04	5052	1
Intercept[5]	0.99	0.10	0.79	1.2	5100	1
Intercept[6]	2.44	0.11	2.23	2.65	5455	1
MusSoph	-0.16	0.10	-0.35	0.03	4297	1
Height	0.13	0.08	-0.03	0.28	4850	1
RelHeight	0.34	0.08	0.18	0.49	4700	1
Height <sup>2</sup>	0.06	0.04	-0.02	0.15	7863	1
RelHeight <sup>2</sup>	-0.31	0.06	-0.43	-0.2	6595	1
Primacy	-0.24	0.12	-0.48	-0.01	13697	1
Previous	0.33	0.07	0.19	0.47	5923	1
Recency	0.54	0.14	0.26	0.82	11496	1
SPCS	0.48	0.09	0.30	0.65	5687	1
MelCont	-0.04	0.03	-0.11	0.02	16017	1
Count	0.01	0.05	-0.09	0.12	11844	1
TrialNo	0.05	0.05	-0.04	0.15	11322	1
InContTrialNo	-0.05	0.04	-0.12	0.02	16338	1
ITrialNo <sup>2</sup>	0.02	0.04	-0.07	0.11	12731	1
ScaleSize	-0.02	0.05	-0.11	0.06	11648	1
BlockOrder	0.00	0.19	-0.38	0.37	3764	1
MusSoph:Height	0.12	0.07	-0.03	0.27	4773	1
MusSoph:RelHeight	0.00	0.08	-0.15	0.15	4491	1
MusSoph:Height <sup>2</sup>	0.01	0.04	-0.07	0.09	7270	1
MusSoph:RelHeight <sup>2</sup>	0.05	0.05	-0.05	0.16	5925	1
MusSoph:Primacy	-0.06	0.09	-0.23	0.11	16561	1
MusSoph:Previous	-0.09	0.07	-0.22	0.04	5458	1
MusSoph:Recency	0.32	0.12	0.09	0.57	11244	1
MusSoph:SPCS	0.29	0.09	0.11	0.46	5014	1
MusSoph:MelCont	0.01	0.02	-0.04	0.05	21120	1
MusSoph:Count	-0.01	0.04	-0.09	0.06	11092	1
Height:TrialNo	0.03	0.04	-0.06	0.11	10291	1
RelHeight:TrialNo	-0.02	0.04	-0.1	0.06	11802	1
Height <sup>2</sup> :TrialNo	-0.04	0.02	-0.08	0.01	14407	1
RelHeight <sup>2</sup> :TrialNo	0.04	0.03	-0.01	0.09	16471	1
Primacy:TrialNo	0.05	0.09	-0.12	0.22	15413	1
Previous:TrialNo	-0.01	0.03	-0.07	0.04	13122	1
Recency:TrialNo	-0.01	0.09	-0.2	0.17	17821	1
SPCS:TrialNo	-0.05	0.03	-0.1	0.01	14974	1
MelCont:TrialNo	-0.06	0.02	-0.11	-0.02	17790	1
Count:TrialNo	0.00	0.03	-0.06	0.06	14036	1
Height:InContTrialNo	-0.07	0.03	-0.13	0.00	13908	1
RelHeight:InContTrialNo	0.05	0.03	-0.01	0.11	14381	1
Height <sup>2</sup> :InContTrialNo	-0.02	0.02	-0.07	0.02	16156	1

*Continued on next page*



Table B.6 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
RelHeight <sup>2</sup> :InContTrialNo	0.02	0.03	-0.03	0.08	15137	1
Primacy:InContTrialNo	0.04	0.09	-0.13	0.21	17283	1
Previous:InContTrialNo	-0.00	0.03	-0.05	0.05	16229	1
Recency:InContTrialNo	0.00	0.09	-0.17	0.17	18267	1
SPCS:InContTrialNo	0.01	0.02	-0.03	0.06	17627	1
MelCont:InContTrialNo	-0.02	0.02	-0.07	0.02	20052	1
Count:InContTrialNo	-0.02	0.03	-0.07	0.03	16062	1
Height:TrialsNo <sup>2</sup>	0.03	0.04	-0.04	0.10	12078	1
RelHeight:TrialsNo <sup>2</sup>	0.03	0.03	-0.04	0.10	14274	1
Height <sup>2</sup> :TrialsNo <sup>2</sup>	0.01	0.02	-0.04	0.06	12688	1
RelHeight <sup>2</sup> :TrialsNo <sup>2</sup>	-0.01	0.03	-0.07	0.04	14222	1
Primacy:TrialsNo <sup>2</sup>	0.06	0.09	-0.12	0.25	12258	1
Previous:TrialsNo <sup>2</sup>	-0.05	0.03	-0.11	0.02	11656	1
Recency:TrialsNo <sup>2</sup>	-0.06	0.10	-0.25	0.13	14429	1
SPCS:TrialsNo <sup>2</sup>	-0.03	0.03	-0.09	0.03	13307	1
MelCont:TrialsNo <sup>2</sup>	0.04	0.02	-0.01	0.09	14209	1
Count:TrialsNo <sup>2</sup>	-0.01	0.03	-0.07	0.05	12868	1
Height:ScaleSize	0.01	0.03	-0.06	0.07	14830	1
RelHeight:ScaleSize	0.05	0.03	-0.01	0.11	15486	1
Height <sup>2</sup> :ScaleSize	0.05	0.02	0.00	0.10	15589	1
RelHeight:ScaleSize	-0.02	0.03	-0.08	0.03	16247	1
Primacy:ScaleSize	-0.01	0.10	-0.21	0.19	12203	1
Previous:ScaleSize	0.00	0.03	-0.05	0.06	14308	1
Recency:ScaleSize	0.07	0.09	-0.1	0.24	17445	1
SPCS:ScaleSize	-0.09	0.03	-0.15	-0.04	16106	1
MelCont:ScaleSize	0.01	0.02	-0.03	0.06	17771	1
Count:ScaleSize	0.02	0.03	-0.04	0.07	12209	1
Height:BlockOrder	-0.13	0.15	-0.42	0.16	4434	1
RelHeight:BlockOrder	0.07	0.15	-0.22	0.36	4457	1
Height <sup>2</sup> :BlockOrder	0.06	0.08	-0.1	0.21	7320	1
RelHeight <sup>2</sup> :BlockOrder	0.07	0.11	-0.14	0.27	5822	1
Primacy:BlockOrder	-0.14	0.17	-0.47	0.19	17892	1
Previous:BlockOrder	-0.01	0.13	-0.27	0.24	6073	1
Recency:BlockOrder	-0.31	0.24	-0.76	0.16	10459	1
SPCS:BlockOrder	0.07	0.17	-0.26	0.41	5064	1
MelCont:BlockOrder	-0.02	0.04	-0.11	0.07	20782	1
Count:BlockOrder	0.00	0.09	-0.17	0.17	12612	1

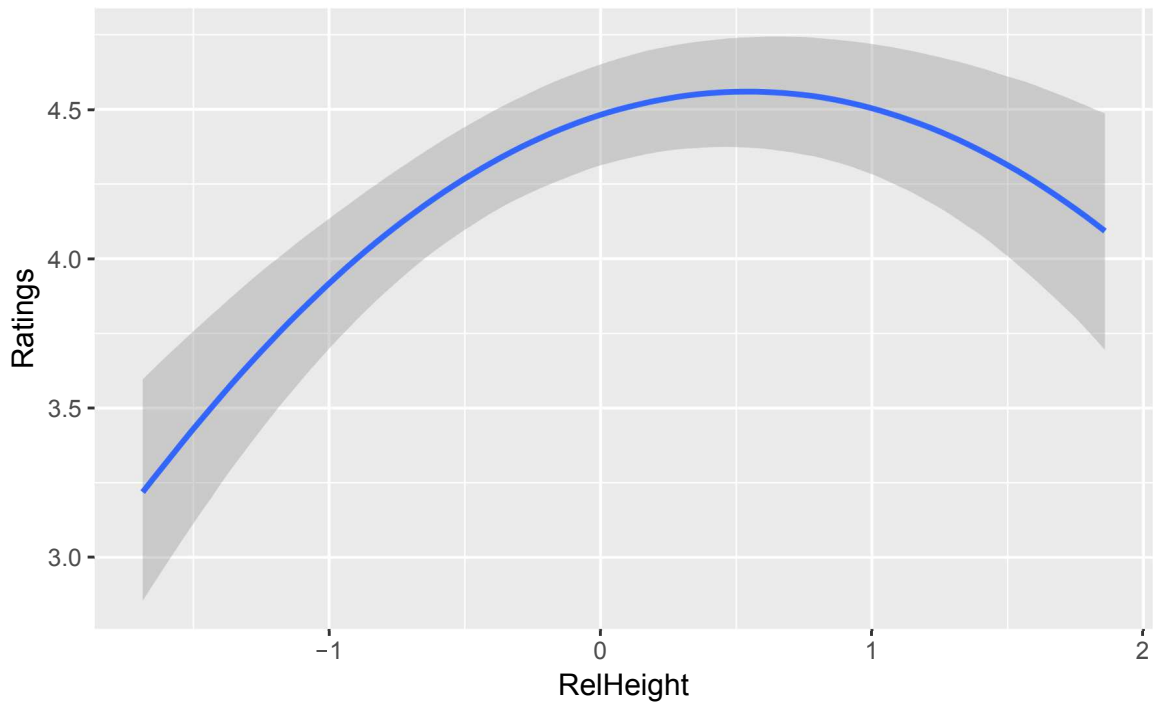


FIGURE B.3: Conditional effect of RelHeight for Model 2.2

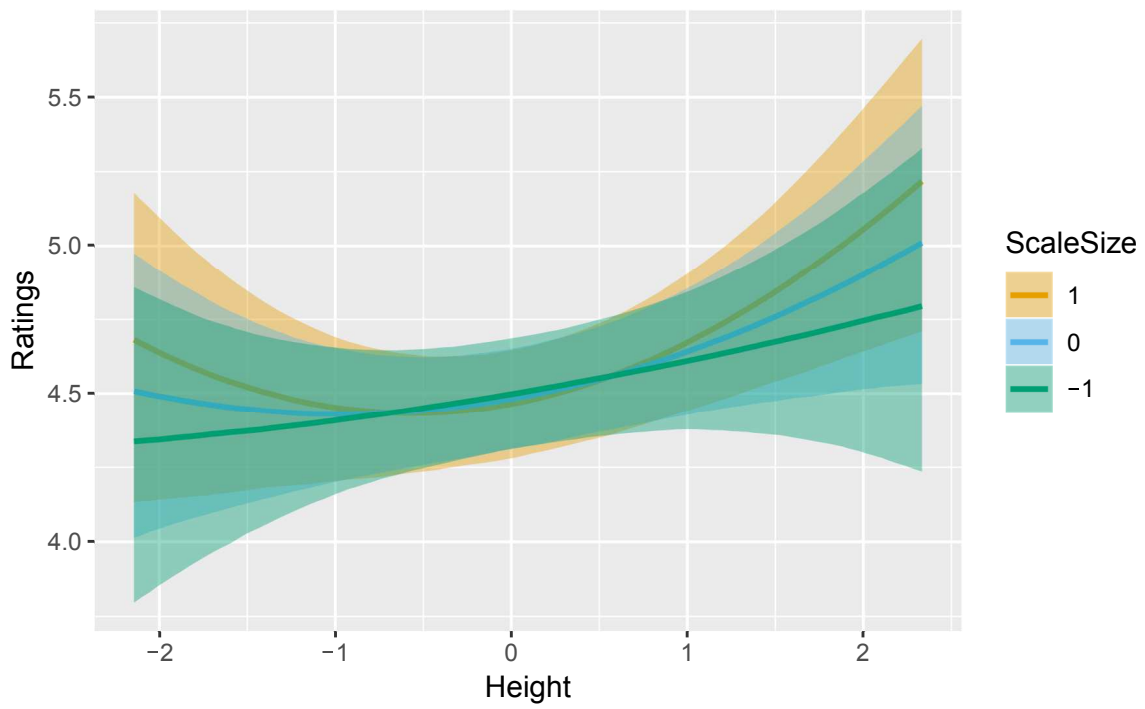


FIGURE B.4: Conditional effect of Height:ScaleSize for Model 2.2

TABLE B.7: All population-level effects for Model 2.3, which was summarized in Table 2.6

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept <sup>[1]</sup>	-3.08	0.14	-3.35	-2.82	1308	1
Intercept <sup>[2]</sup>	-1.38	0.13	-1.64	-1.13	1373	1
Intercept <sup>[3]</sup>	-0.09	0.13	-0.33	0.16	1372	1
Intercept <sup>[4]</sup>	0.66	0.13	0.42	0.91	1350	1
Intercept <sup>[5]</sup>	1.92	0.13	1.65	2.17	1486	1
Intercept <sup>[6]</sup>	3.46	0.14	3.19	3.74	1544	1
MusSoph	-0.12	0.10	-0.32	0.08	682	1
Height	0.15	0.09	-0.02	0.32	1364	1
RelHeight	0.26	0.07	0.12	0.40	1682	1
Height <sup>2</sup>	0.00	0.06	-0.12	0.11	1194	1
RelHeight <sup>2</sup>	-0.11	0.06	-0.22	-0.01	1896	1
Primacy	0.10	0.17	-0.25	0.43	2770	1
Previous	0.32	0.06	0.19	0.45	1717	1
Recency	0.61	0.26	0.12	1.14	1306	1
SPCS	0.88	0.09	0.70	1.07	1425	1
MelCont	-0.03	0.05	-0.13	0.06	2301	1
Count	-0.05	0.10	-0.25	0.15	1595	1
TrialNo	-0.01	0.06	-0.14	0.11	1715	1
InContTrialNo	0.04	0.06	-0.07	0.15	2517	1
TrialNo <sup>2</sup>	0.06	0.06	-0.07	0.18	2180	1
BlockOrder	0.04	0.22	-0.4	0.45	1070	1
MusSoph:Height	0.01	0.07	-0.13	0.14	1347	1
MusSoph:RelHeight	-0.09	0.05	-0.2	0.01	1870	1
MusSoph:Height <sup>2</sup>	0.01	0.05	-0.09	0.11	1183	1.01
MusSoph:RelHeight <sup>2</sup>	-0.01	0.04	-0.1	0.07	2502	1
MusSoph:Primacy	-0.21	0.13	-0.47	0.04	2998	1
MusSoph:Previous	-0.12	0.05	-0.23	-0.02	1439	1
MusSoph:Recency	0.59	0.23	0.15	1.04	1381	1
MusSoph:SPCS	0.41	0.09	0.23	0.59	1190	1
MusSoph:MelCont	0.07	0.04	-0.01	0.14	2889	1
MusSoph:Count	0.03	0.05	-0.07	0.14	2183	1
Height:TrialNo	0.04	0.05	-0.05	0.13	2474	1
RelHeight:TrialNo	0.00	0.05	-0.09	0.09	2615	1
Height <sup>2</sup> :TrialNo	0.07	0.03	0.00	0.14	3063	1
RelHeight <sup>2</sup> :TrialNo	0.01	0.04	-0.06	0.09	2570	1
Primacy:TrialNo	-0.01	0.14	-0.27	0.26	2664	1
Previous:TrialNo	0.04	0.04	-0.03	0.11	2375	1
Recency:TrialNo	0.24	0.14	-0.03	0.51	2914	1
SPCS:TrialNo	-0.08	0.04	-0.16	-0.01	2208	1
MelCont:TrialNo	-0.04	0.04	-0.12	0.05	1680	1
Count:TrialNo	0.03	0.04	-0.05	0.11	2230	1
Height:InContTrialNo	-0.07	0.05	-0.16	0.03	2361	1
RelHeight:InContTrialNo	0.07	0.04	-0.02	0.16	2203	1
Height <sup>2</sup> :InContTrialNo	-0.07	0.03	-0.13	0.00	2327	1
RelHeight <sup>2</sup> :InContTrialNo	0.00	0.04	-0.07	0.09	2720	1
Primacy:InContTrialNo	0.11	0.14	-0.15	0.37	2258	1
Previous:InContTrialNo	0.05	0.04	-0.03	0.13	2094	1
Recency:InContTrialNo	0.07	0.13	-0.21	0.32	3174	1
SPCS:InContTrialNo	0.08	0.04	0.00	0.16	2163	1
MelCont:InContTrialNo	0.01	0.04	-0.07	0.08	3071	1

*Continued on next page*

Table B.7 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Count:lnContTrialNo	-0.03	0.04	-0.11	0.05	2073	1
Height: TrialNo <sup>2</sup>	-0.06	0.05	-0.16	0.04	1966	1
RelHeight: TrialNo <sup>2</sup>	-0.02	0.05	-0.11	0.08	1823	1
Height <sup>2</sup> : TrialNo <sup>2</sup>	-0.05	0.04	-0.12	0.01	1802	1
RelHeight <sup>2</sup> : TrialNo <sup>2</sup>	-0.01	0.04	-0.09	0.08	2099	1
Primacy: TrialNo <sup>2</sup>	0.00	0.14	-0.27	0.28	2448	1
Previous: TrialNo <sup>2</sup>	-0.04	0.04	-0.11	0.04	1869	1
Recency: TrialNo <sup>2</sup>	0.09	0.15	-0.21	0.38	2254	1
SPCS: TrialNo <sup>2</sup>	0.01	0.05	-0.08	0.11	2098	1
MelCont: TrialNo <sup>2</sup>	0.04	0.04	-0.03	0.12	2093	1
Count: TrialNo <sup>2</sup>	-0.02	0.04	-0.11	0.06	1697	1
Height: BlockOrder	-0.02	0.13	-0.28	0.24	1253	1
RelHeight: BlockOrder	0.04	0.10	-0.16	0.25	2524	1
Height <sup>2</sup> : BlockOrder	0.02	0.10	-0.18	0.21	1497	1
RelHeight <sup>2</sup> : BlockOrder	0.10	0.08	-0.06	0.25	2454	1
Primacy: BlockOrder	0.00	0.25	-0.48	0.50	2337	1
Previous: BlockOrder	-0.06	0.11	-0.27	0.15	1578	1
Recency: BlockOrder	0.29	0.44	-0.56	1.13	1620	1
SPCS: BlockOrder	-0.32	0.18	-0.69	0.02	1163	1
MelCont: BlockOrder	0.07	0.07	-0.08	0.22	2039	1
Count: BlockOrder	0.14	0.19	-0.24	0.51	1883	1
Count: BlockOrder	0.14	0.19	-0.24	0.51	1883	1
Height: BlockOrder	-0.13	0.15	-0.42	0.16	4434	1
RelHeight: BlockOrder	0.07	0.15	-0.22	0.36	4457	1
Height <sup>2</sup> : BlockOrder	0.06	0.08	-0.1	0.21	7320	1
RelHeight <sup>2</sup> : BlockOrder	0.07	0.11	-0.14	0.27	5822	1
Primacy: BlockOrder	-0.14	0.17	-0.47	0.19	17892	1
Previous: BlockOrder	-0.01	0.13	-0.27	0.24	6073	1
Recency: BlockOrder	-0.31	0.24	-0.76	0.16	10459	1
SPCS: BlockOrder	0.07	0.17	-0.26	0.41	5064	1
MelCont: BlockOrder	-0.02	0.04	-0.11	0.07	20782	1
Count: BlockOrder	0.00	0.09	-0.17	0.17	12612	1

## **Appendix C**

### **Experiments 1 & 2 Triads**

#### **C.1 Experiment 1**

##### **C.1.1 Pre-registered model**

TABLE C.1: Pre-registered probe triads model significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.16	0.22	-3.57	-2.73	564	1.00
Intercept[2]	-1.75	0.21	-2.15	-1.32	558	1.00
Intercept[3]	-0.57	0.21	-0.98	-0.13	566	1.00
Intercept[4]	0.08	0.21	-0.33	0.51	566	1.00
Intercept[5]	1.24	0.21	0.83	1.67	567	1.00
Intercept[6]	2.74	0.21	2.33	3.17	590	1.00
MusSoph	-0.25	0.13	-0.52	-0.00	770	1.01
Recency	0.26	0.10	0.06	0.46	718	1.00
SPCS	0.34	0.09	0.16	0.50	752	1.00
Count	-0.03	0.07	-0.17	0.11	1349	1.00
Minor	-0.55	0.17	-0.89	-0.20	544	1.00
Diminished	-0.79	0.23	-1.23	-0.34	572	1.00
Augmented	-1.04	0.29	-1.62	-0.42	551	1.00
Prevalence	0.32	0.09	0.14	0.50	842	1.00
InContTrialNo	-0.00	0.09	-0.17	0.17	1010	1.00
BlockOrder	-0.12	0.26	-0.63	0.40	613	1.01
MusSoph:SPCS	0.23	0.06	0.11	0.35	787	1.00
MusSoph:Minor	0.26	0.11	0.03	0.49	708	1.00
MusSoph:Diminished	0.52	0.15	0.22	0.80	860	1.00
MusSoph:Augmented	0.15	0.19	-0.23	0.51	885	1.00
MusSoph:Prevalence	0.25	0.07	0.11	0.37	960	1.01
Count:InContTrialNo	-0.09	0.03	-0.14	-0.03	2191	1.00
Count:BlockOrder	-0.29	0.10	-0.48	-0.10	1615	1.01

Significant population-level effects for the preregistered probe triads model along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ )’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

A LOO comparison was made between this model and one identical to it but for the absence of SPCS and Prevalence as predictors. The preregistered model is found to significantly outperform the other, with a difference in LOOIC of 903.6, for a SE of 74.8, confirming H2.

Given that interaction effects with MusSoph are significant (MusSoph:SPCS, MusSoph:Prevalence, and interactions of MusSoph with Triad), H3 is also confirmed.

### C.1.2 Model 3.1

TABLE C.2: All population-level effects for Model 3.1, which was summarized in Table 3.2

Effect	Estimate	Est. Error	l-95% CI	u-95% CI	Eff. Sample	Rhat
Intercept[1]	-3.24	0.14	-3.51	-2.97	5072	1.00
Intercept[2]	-1.77	0.13	-2.03	-1.51	5053	1.00
Intercept[3]	-0.54	0.13	-0.80	-0.29	5003	1.00
Intercept[4]	0.12	0.13	-0.13	0.38	5007	1.00
Intercept[5]	1.33	0.13	1.08	1.59	5107	1.00
Intercept[6]	2.89	0.14	2.62	3.16	5311	1.00
MusSoph	-0.11	0.11	-0.33	0.10	3777	1.00
Height	0.08	0.25	-0.42	0.57	5451	1.00
RelHeight	-0.11	0.25	-0.60	0.38	5393	1.00
Height <sup>2</sup>	-0.08	0.12	-0.31	0.16	5538	1.00
RelHeight <sup>2</sup>	0.09	0.12	-0.14	0.32	5800	1.00
Primacy	0.02	0.06	-0.11	0.15	11466	1.00
Previous	0.24	0.05	0.14	0.33	6619	1.00
Recency	0.24	0.09	0.08	0.41	7526	1.00
SPCS	0.39	0.07	0.26	0.53	4852	1.00
MelCont	0.05	0.03	-0.02	0.11	10398	1.00
Count	0.03	0.06	-0.09	0.15	9259	1.00
Minor	-0.43	0.13	-0.69	-0.18	4815	1.00
Diminished	-0.59	0.17	-0.92	-0.25	5063	1.00
Augmented	-1.06	0.21	-1.47	-0.65	5611	1.00
Prevalence	0.32	0.07	0.18	0.46	5939	1.00
TrialNo	0.09	0.07	-0.05	0.23	6101	1.00
InContTrialNo	-0.08	0.06	-0.20	0.04	5485	1.00
TrialNo <sup>2</sup>	0.10	0.07	-0.05	0.25	5927	1.00
Order	-0.04	0.23	-0.50	0.42	3272	1.00
MusSoph:Height	-0.04	0.21	-0.44	0.37	6903	1.00
MusSoph:RelHeight	0.07	0.20	-0.33	0.47	6902	1.00
MusSoph:Height <sup>2</sup>	0.07	0.10	-0.12	0.27	6261	1.00
MusSoph:RelHeight <sup>2</sup>	-0.08	0.10	-0.27	0.11	6130	1.00
MusSoph:Primacy	-0.07	0.05	-0.17	0.03	12891	1.00
MusSoph:Previous	0.04	0.04	-0.05	0.12	5912	1.00
MusSoph:Recency	0.06	0.07	-0.08	0.20	6497	1.00
MusSoph:SPCS	0.22	0.06	0.10	0.35	5311	1.00
MusSoph:MelCont	0.01	0.02	-0.04	0.05	17248	1.00
MusSoph:Count	0.06	0.04	-0.01	0.14	8710	1.00
MusSoph:Minor	0.26	0.11	0.03	0.48	4749	1.00
MusSoph:Diminished	0.52	0.15	0.22	0.82	5122	1.00
MusSoph:Augmented	0.14	0.18	-0.22	0.50	5313	1.00
MusSoph:Prevalence	0.24	0.06	0.12	0.37	6540	1.00
Height:TrialNo	0.45	0.24	-0.01	0.92	4823	1.00
RelHeight:TrialNo	-0.47	0.24	-0.95	-0.02	4723	1.00
Height <sup>2</sup> :TrialNo	-0.02	0.11	-0.23	0.20	6014	1.00
RelHeight <sup>2</sup> :TrialNo	-0.00	0.11	-0.22	0.21	6203	1.00
Primacy:TrialNo	-0.01	0.05	-0.11	0.08	14425	1.00
Previous:TrialNo	-0.02	0.03	-0.07	0.04	10702	1.00
Recency:TrialNo	0.03	0.05	-0.07	0.12	15495	1.00
SPCS:TrialNo	0.04	0.03	-0.03	0.10	7569	1.00
MelCont:TrialNo	-0.02	0.03	-0.08	0.03	11022	1.00
Count:TrialNo	-0.02	0.05	-0.12	0.07	7224	1.00
Minor:TrialNo	-0.02	0.07	-0.15	0.11	6500	1.00

*Continued on next page*

Table C.2 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Diminished: TrialNo	-0.07	0.09	-0.25	0.10	5912	1.00
Augmented: TrialNo	-0.00	0.13	-0.26	0.26	7165	1.00
Prevalence: TrialNo	-0.00	0.05	-0.10	0.10	6205	1.00
Height: lnContTrialNo	-0.06	0.15	-0.37	0.24	7303	1.00
RelHeight: lnContTrialNo	0.04	0.16	-0.27	0.34	7207	1.00
Height <sup>2</sup> : lnContTrialNo	0.05	0.08	-0.10	0.20	8068	1.00
RelHeight <sup>2</sup> : lnContTrialNo	-0.04	0.08	-0.19	0.11	8119	1.00
Primacy: lnCont TrialNo	0.04	0.05	-0.06	0.13	13604	1.00
Previous: lnCont TrialNo	0.03	0.03	-0.03	0.08	12208	1.00
Recency: lnCont TrialNo	-0.01	0.05	-0.11	0.08	12931	1.00
SPCS: lnCont TrialNo	-0.02	0.03	-0.07	0.03	10397	1.00
MelCont: lnCont TrialNo	0.00	0.03	-0.05	0.05	11639	1.00
Count: lnCont TrialNo	-0.08	0.03	-0.14	-0.03	11670	1.00
Minor: lnCont TrialNo	0.12	0.06	-0.01	0.24	6250	1.00
Diminished: lnCont TrialNo	0.14	0.08	-0.03	0.31	5370	1.00
Augmented: lnCont TrialNo	0.20	0.11	-0.01	0.42	7020	1.00
Prevalence: lnCont TrialNo	0.01	0.04	-0.06	0.08	5886	1.00
Height: TrialNo <sup>2</sup>	0.17	0.22	-0.26	0.59	4411	1.00
RelHeight: TrialNo <sup>2</sup>	-0.18	0.22	-0.60	0.24	4340	1.00
Height <sup>2</sup> : TrialNo <sup>2</sup>	0.04	0.10	-0.16	0.23	5708	1.00
RelHeight <sup>2</sup> : TrialNo <sup>2</sup>	-0.05	0.10	-0.24	0.15	5705	1.00
Primacy: TrialNo <sup>2</sup>	-0.04	0.05	-0.15	0.07	10239	1.00
Previous: TrialNo <sup>2</sup>	0.02	0.03	-0.04	0.07	11488	1.00
Recency: TrialNo <sup>2</sup>	0.02	0.05	-0.08	0.12	10501	1.00
SPCS: TrialNo <sup>2</sup>	0.03	0.04	-0.04	0.10	7630	1.00
MelCont: TrialNo <sup>2</sup>	0.03	0.02	-0.02	0.07	10061	1.00
Count: TrialNo <sup>2</sup>	-0.06	0.03	-0.12	0.00	9983	1.00
Minor: TrialNo <sup>2</sup>	0.03	0.07	-0.11	0.17	5722	1.00
Diminished: TrialNo <sup>2</sup>	-0.06	0.10	-0.25	0.14	5263	1.00
Augmented: TrialNo <sup>2</sup>	-0.07	0.14	-0.33	0.20	6214	1.00
Prevalence: TrialNo <sup>2</sup>	0.02	0.05	-0.08	0.12	5833	1.00
Height: BlockOrder	0.06	0.40	-0.72	0.84	6305	1.00
RelHeight: BlockOrder	0.00	0.39	-0.77	0.77	6259	1.00
Height <sup>2</sup> : BlockOrder	0.08	0.19	-0.29	0.45	7368	1.00
RelHeight <sup>2</sup> : BlockOrder	-0.05	0.19	-0.42	0.31	7587	1.00
Primacy: BlockOrder	-0.03	0.10	-0.23	0.16	14353	1.00
Previous: BlockOrder	0.14	0.09	-0.04	0.32	5292	1.00
Recency: BlockOrder	0.13	0.15	-0.16	0.43	6147	1.00
SPCS: BlockOrder	0.01	0.13	-0.25	0.26	4714	1.00
MelCont: BlockOrder	0.03	0.04	-0.06	0.11	14952	1.00
Count: BlockOrder	-0.30	0.10	-0.49	-0.11	9238	1.00
Minor: BlockOrder	-0.07	0.23	-0.53	0.38	4704	1.00
Diminished: BlockOrder	-0.15	0.31	-0.76	0.46	4769	1.00
Augmented: BlockOrder	-0.50	0.37	-1.22	0.23	4970	1.00
Prevalence: BlockOrder	-0.14	0.13	-0.40	0.12	5973	1.00



TABLE C.3: Model 3.1 but without Prevalence significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.69	0.15	-3.98	-3.40	638	1.00
Intercept[2]	-2.23	0.14	-2.49	-1.96	635	1.00
Intercept[3]	-1.02	0.14	-1.29	-0.75	640	1.00
Intercept[4]	-0.37	0.14	-0.64	-0.09	648	1.00
Intercept[5]	0.81	0.14	0.54	1.08	634	1.00
Intercept[6]	2.30	0.14	2.02	2.58	667	1.00
MusSoph	0.05	0.12	-0.18	0.31	437	1.00
Previous	0.24	0.05	0.14	0.34	842	1.00
Recency	-0.23	0.09	-0.40	-0.06	1051	1.00
SPCS	0.50	0.07	0.37	0.63	769	1.00
Count	0.00	0.06	-0.12	0.13	1225	1.00
Minor	-0.70	0.13	-0.95	-0.46	701	1.00
Diminished	-1.09	0.17	-1.41	-0.75	634	1.00
Augmented	-1.57	0.21	-2.00	-1.18	731	1.00
TrialNo	0.17	0.07	0.04	0.30	1283	1.00
InContTrialNo	-0.07	0.06	-0.18	0.05	990	1.01
BlockOrder	-0.12	0.25	-0.64	0.36	448	1.01
MusSoph:SPCS	0.29	0.06	0.17	0.41	590	1.00
MusSoph:Count	0.08	0.04	0.00	0.16	1214	1.00
Minor:TrialNo	-0.11	0.06	-0.22	-0.00	1393	1.00
Diminished:TrialNo	-0.13	0.06	-0.25	-0.02	1534	1.00
Augmented:TrialNo	-0.04	0.10	-0.24	0.18	1621	1.00
Count:InContTrialNo	-0.08	0.03	-0.14	-0.03	1485	1.00
Minor:InContTrialNo	0.13	0.05	0.04	0.23	1722	1.00
Diminished:InContTrialNo	0.14	0.05	0.03	0.25	1487	1.00
Augmented:InContTrialNo	0.18	0.09	0.01	0.36	2104	1.00
Count:BlockOrder	-0.30	0.10	-0.50	-0.12	1134	1.00

Significant population-level effects for Model 3.1 but without Prevalence along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ‘:’ written between the interacting effects. As written in *brms*, ‘for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)’. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

TABLE C.4: Model 3.1 but with ScaleTone instead of SPCS significant population-level effects

Intercept[1]	-3.44	0.15	-3.73	-3.16	576	1.01
Intercept[2]	-1.97	0.14	-2.24	-1.68	616	1.00
Intercept[3]	-0.75	0.14	-1.02	-0.48	607	1.00
Intercept[4]	-0.08	0.14	-0.35	0.20	613	1.00
Intercept[5]	1.13	0.14	0.86	1.40	611	1.00
Intercept[6]	2.68	0.14	2.41	2.97	650	1.00
MusSoph	-0.14	0.12	-0.37	0.07	562	1.01
Height	0.12	0.26	-0.38	0.61	665	1.00
RelHeight	-0.16	0.25	-0.67	0.33	643	1.00
Previous	0.23	0.05	0.14	0.33	832	1.00
Recency	-0.24	0.08	-0.40	-0.08	907	1.00
ScaleTone	0.36	0.07	0.24	0.50	692	1.00
Count	0.03	0.06	-0.08	0.15	1132	1.00
Minor	-0.35	0.13	-0.61	-0.10	658	1.01
Diminished	-0.47	0.16	-0.79	-0.14	778	1.00
Augmented	-1.00	0.21	-1.41	-0.59	764	1.01
Prevalence	0.37	0.07	0.23	0.51	890	1.00
TrialNo	0.10	0.08	-0.05	0.26	1075	1.00
InContTrialNo	-0.06	0.07	-0.20	0.07	922	1.01
BlockOrder	0.04	0.25	-0.47	0.54	482	1.01
MusSoph:ScaleTone	0.21	0.06	0.09	0.35	723	1.00
MusSoph:Minor	0.29	0.11	0.07	0.52	661	1.01
MusSoph:Diminished	0.59	0.16	0.29	0.90	687	1.01
MusSoph:Augmented	0.19	0.19	-0.17	0.57	582	1.00
MusSoph:Prevalence	0.27	0.07	0.14	0.40	586	1.00
Height:TrialNo	0.51	0.23	0.06	0.97	686	1.01
RelHeight:TrialNo	-0.54	0.23	-1.00	-0.08	684	1.01
Count:InContTrialNo	-0.08	0.03	-0.13	-0.03	2101	1.00
Count:BlockOrder	-0.32	0.10	-0.52	-0.12	1672	1.00

Significant population-level effects for Model 3.1 but with ScaleTone instead of SPCS along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

TABLE C.5: Model 3.1 but without Prevalence and with ScaleTone instead of SPCS significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.65	0.14	-3.93	-3.36	449	1.01
Intercept[2]	-2.20	0.14	-2.48	-1.93	456	1.01
Intercept[3]	-1.00	0.14	-1.27	-0.73	459	1.01
Intercept[4]	-0.35	0.14	-0.62	-0.07	457	1.01
Intercept[5]	0.82	0.14	0.55	1.10	458	1.01
Intercept[6]	2.31	0.14	2.03	2.60	476	1.01
MusSoph	0.04	0.11	-0.20	0.27	436	1.01
Previous	0.24	0.05	0.15	0.34	627	1.00
Recency	-0.23	0.09	-0.40	-0.06	845	1.00
ScaleTone	0.47	0.07	0.34	0.61	674	1.00
Count	0.04	0.06	-0.09	0.15	993	1.01
Minor	-0.67	0.12	-0.93	-0.45	446	1.02
Diminished	-1.04	0.17	-1.37	-0.70	476	1.00
Augmented	-1.60	0.21	-2.00	-1.20	612	1.00
TrialNo	0.17	0.07	0.04	0.30	1095	1.00
InContTrialNo	-0.09	0.06	-0.21	0.03	845	1.00
BlockOrder	-0.04	0.24	-0.55	0.41	376	1.01
MusSoph:ScaleTone	0.26	0.06	0.13	0.38	685	1.00
MusSoph:Count	0.09	0.04	0.01	0.17	1272	1.00
Minor:TrialNo	-0.14	0.06	-0.25	-0.02	1117	1.00
Diminished:TrialNo	-0.14	0.06	-0.25	-0.02	1491	1.00
Augmented:TrialNo	-0.06	0.10	-0.27	0.13	1497	1.00
Count:InContTrialNo	-0.08	0.03	-0.13	-0.03	1519	1.00
Minor:InContTrialNo	0.16	0.05	0.07	0.26	1351	1.00
Diminished:InContTrialNo	0.15	0.06	0.04	0.26	1405	1.00
Augmented:InContTrialNo	0.20	0.09	0.03	0.37	1432	1.00
MelCont:BlockOrder	0.02	0.05	-0.07	0.11	2656	1.00
Count:BlockOrder	-0.32	0.10	-0.52	-0.12	1052	1.00

Significant population-level effects for Model 3.1 but without Prevalence and with ScaleTone instead of SPCS along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ ). The chains are obtained via *Markov chain Monte Carlo* (MCMC), which are a class of algorithms used in *brms* for sampling from a probability distribution.

## C.2 Experiment 2

### C.2.1 Model 3.2

TABLE C.6: All population-level effects for Model 3.2, which was summarized in Table 3.6

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.53	0.19	-3.90	-3.16	5837	1
Intercept[2]	-2.34	0.18	-2.70	-1.98	5919	1
Intercept[3]	-1.16	0.18	-1.51	-0.80	5965	1
Intercept[4]	-0.31	0.18	-0.66	0.05	5958	1
Intercept[5]	0.94	0.18	0.59	1.30	6049	1
Intercept[6]	2.46	0.19	2.10	2.83	6293	1
MusSoph	0.38	0.15	0.08	0.68	4300	1
Height	-0.53	0.30	-1.11	0.05	7295	1
RelHeight	0.38	0.29	-0.20	0.95	7586	1
Height <sup>2</sup>	0.00	0.14	-0.26	0.26	6539	1
RelHeight <sup>2</sup>	0.04	0.14	-0.22	0.31	6711	1
Primacy	-0.06	0.09	-0.23	0.11	13451	1
Previous	0.32	0.07	0.18	0.46	8016	1
Recency	-0.06	0.09	-0.23	0.12	13749	1
SPCS	0.27	0.06	0.15	0.38	10689	1
MelCont	0.01	0.04	-0.07	0.10	13356	1
Count	-0.02	0.11	-0.23	0.20	10693	1
Minor	-0.72	0.14	-0.99	-0.46	8748	1
Diminished	-1.18	0.17	-1.51	-0.86	8011	1
Augmented	-1.26	0.20	-1.66	-0.86	8835	1
TrialNo	0.02	0.09	-0.16	0.19	9948	1
IncontTrialNo	-0.08	0.07	-0.22	0.06	11286	1
TrialNo <sup>2</sup>	0.04	0.08	-0.13	0.20	9418	1
BlockOrder	-0.46	0.32	-1.10	0.16	5040	1
MusSoph:Height	-0.29	0.30	-0.89	0.30	6275	1
MusSoph:RelHeight	0.44	0.30	-0.13	1.03	6387	1
MusSoph:Height <sup>2</sup>	0.01	0.11	-0.20	0.22	7982	1
MusSoph:RelHeight <sup>2</sup>	0.00	0.11	-0.22	0.21	7861	1
MusSoph:Primacy	0.03	0.07	-0.10	0.16	16139	1
MusSoph:Previous	-0.10	0.06	-0.23	0.02	7202	1
MusSoph:Recency	-0.04	0.07	-0.16	0.09	17030	1
MusSoph:SPCS	0.15	0.05	0.05	0.24	9910	1
MusSoph:MelCont	0.02	0.03	-0.04	0.08	16348	1
MusSoph:Count	0.02	0.07	-0.11	0.16	10920	1
MusSoph:Minor	-0.38	0.12	-0.60	-0.15	8309	1
MusSoph:Diminished	-0.33	0.15	-0.63	-0.03	6917	1
MusSoph:Augmented	-0.53	0.16	-0.85	-0.21	9291	1
Height:TrialNo	0.01	0.32	-0.62	0.65	5837	1
RelHeight:TrialNo	-0.02	0.33	-0.67	0.61	5820	1
Height <sup>2</sup> :TrialNo	0.12	0.11	-0.10	0.35	7872	1
RelHeight <sup>2</sup> :TrialNo	-0.13	0.11	-0.35	0.08	8076	1
Primacy:TrialNo	0.02	0.06	-0.10	0.14	16738	1
Previous:TrialNo	-0.01	0.03	-0.08	0.06	17529	1
Recency:TrialNo	-0.05	0.07	-0.19	0.08	14251	1
SPCS:TrialNo	-0.04	0.04	-0.11	0.03	13813	1
MelCont:TrialNo	0.02	0.03	-0.05	0.08	17372	1
Count:TrialNo	-0.06	0.06	-0.17	0.05	9435	1
Minor:TrialNo	-0.03	0.08	-0.19	0.12	11603	1
Diminished:TrialNo	-0.14	0.09	-0.32	0.05	11179	1
Augmented:TrialNo	0.12	0.14	-0.15	0.40	12562	1

*Continued on next page*

Table C.6 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Height:IncontTrialNo	-0.02	0.20	-0.41	0.37	7140	1
RelHeight:IncontTrialNo	0.01	0.20	-0.37	0.41	7136	1
Height <sup>2</sup> :IncontTrialNo	-0.06	0.09	-0.25	0.12	8677	1
RelHeight <sup>2</sup> :IncontTrialNo	0.06	0.10	-0.13	0.25	8947	1
Primacy:IncontTrialNo	0.09	0.06	-0.03	0.22	17263	1
Previous:IncontTrialNo	-0.01	0.04	-0.09	0.07	13199	1
Recency:IncontTrialNo	0.00	0.07	-0.13	0.14	14181	1
SPCS:IncontTrialNo	-0.04	0.04	-0.11	0.03	14924	1
MelCont:IncontTrialNo	0.02	0.03	-0.04	0.08	15290	1
Count:IncontTrialNo	0.01	0.04	-0.06	0.09	12225	1
Minor:IncontTrialNo	-0.01	0.07	-0.15	0.14	12266	1
Diminished:IncontTrialNo	0.10	0.08	-0.07	0.26	12728	1
Augmented:IncontTrialNo	-0.09	0.12	-0.32	0.14	12314	1
Height:TrialNo <sup>2</sup>	0.29	0.20	-0.11	0.69	8032	1
RelHeight:TrialNo <sup>2</sup>	-0.27	0.20	-0.67	0.13	7976	1
Height <sup>2</sup> :TrialNo <sup>2</sup>	0.04	0.09	-0.14	0.23	6516	1
RelHeight <sup>2</sup> :TrialNo <sup>2</sup>	-0.08	0.10	-0.27	0.11	6576	1
Primacy:TrialNo <sup>2</sup>	0.00	0.07	-0.13	0.14	11492	1
Previous:TrialNo <sup>2</sup>	0.01	0.04	-0.06	0.08	13616	1
Recency:TrialNo <sup>2</sup>	0.07	0.07	-0.06	0.20	11997	1
SPCS:TrialNo <sup>2</sup>	0.03	0.04	-0.04	0.11	13137	1
MelCont:TrialNo <sup>2</sup>	-0.03	0.03	-0.10	0.04	12562	1
Count:TrialNo <sup>2</sup>	-0.04	0.04	-0.13	0.04	12237	1
Minor:TrialNo <sup>2</sup>	0.03	0.08	-0.14	0.19	10223	1
Diminished:TrialNo <sup>2</sup>	0.12	0.09	-0.05	0.29	11728	1
Augmented:TrialNo <sup>2</sup>	-0.03	0.13	-0.29	0.23	10954	1
Height:BlockOrder	0.06	0.56	-1.05	1.18	5279	1
RelHeight:BlockOrder	-0.01	0.55	-1.11	1.07	5577	1
Height <sup>2</sup> :BlockOrder	-0.10	0.21	-0.51	0.31	7949	1
RelHeight <sup>2</sup> :BlockOrder	0.16	0.21	-0.25	0.58	8126	1
Primacy:BlockOrder	0.05	0.13	-0.20	0.31	14351	1
Previous:BlockOrder	-0.18	0.12	-0.42	0.05	8529	1
Recency:BlockOrder	-0.18	0.13	-0.43	0.07	16528	1
SPCS:BlockOrder	-0.04	0.09	-0.22	0.15	10385	1
MelCont:BlockOrder	0.05	0.06	-0.07	0.17	18335	1
Count:BlockOrder	-0.13	0.19	-0.51	0.26	11758	1
Minor:BlockOrder	0.46	0.22	0.03	0.90	8552	1
Diminished:BlockOrder	0.45	0.29	-0.13	1.02	7464	1
Augmented:BlockOrder	0.60	0.33	-0.04	1.27	9491	1

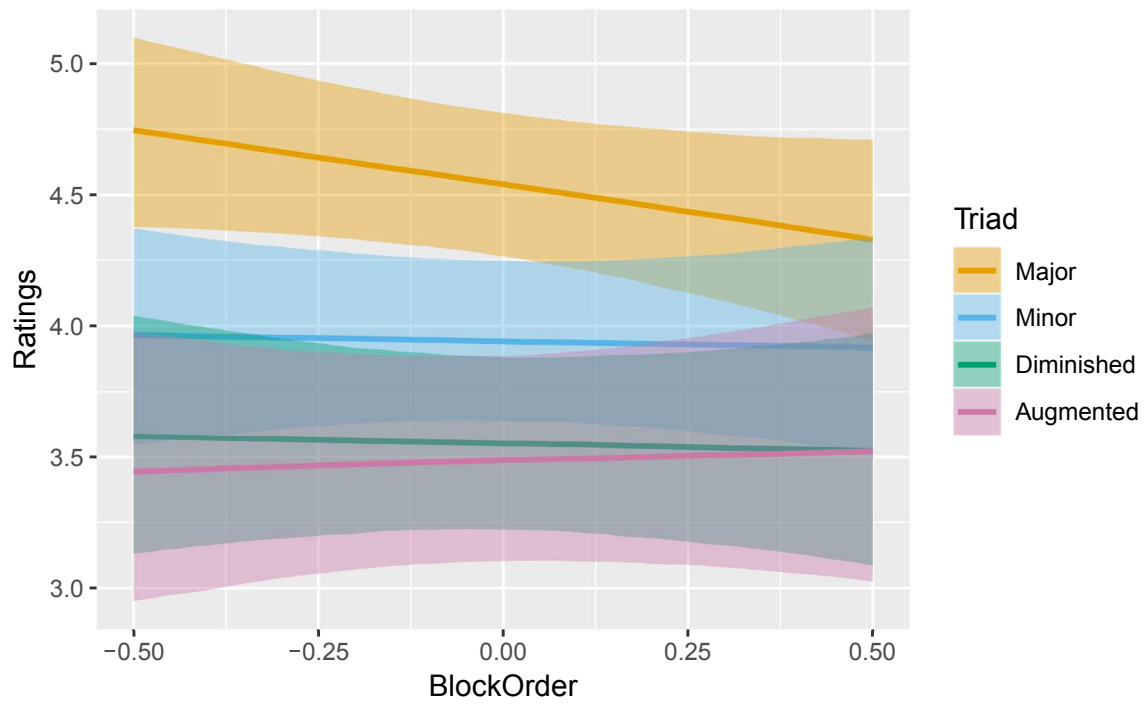


FIGURE C.1: Conditional effect of BlockOrder:Triad for Model 3.2

TABLE C.7: Model 3.2 but with ScaleTone instead of SPCS significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.48	0.19	-3.84	-3.12	1180	1.00
Intercept[2]	-2.28	0.18	-2.64	-1.94	1171	1.00
Intercept[3]	-1.10	0.18	-1.45	-0.76	1157	1.00
Intercept[4]	-0.25	0.18	-0.60	0.10	1149	1.00
Intercept[5]	0.99	0.18	0.63	1.34	1170	1.00
Intercept[6]	2.50	0.18	2.15	2.85	1200	1.00
MusSoph	0.39	0.15	0.09	0.68	795	1.01
Previous	0.32	0.07	0.19	0.46	1485	1.00
ScaleTone	0.24	0.06	0.12	0.37	1482	1.00
Minor	-0.75	0.13	-1.01	-0.50	1529	1.00
Diminished	-1.26	0.16	-1.58	-0.94	1440	1.00
Augmented	-1.22	0.20	-1.61	-0.83	1925	1.00
BlockOrder	-0.53	0.34	-1.18	0.12	867	1.00
MusSoph:ScaleTone	0.15	0.05	0.05	0.25	1770	1.00
MusSoph:Count	0.00	0.08	-0.14	0.15	1715	1.00
MusSoph:Minor	-0.38	0.11	-0.60	-0.16	1716	1.00
MusSoph:Diminished	-0.37	0.14	-0.64	-0.07	1464	1.00
MusSoph:Augmented	-0.49	0.16	-0.81	-0.18	1828	1.00
Minor:BlockOrder	0.47	0.22	0.03	0.90	1435	1.00
Diminished:BlockOrder	0.49	0.29	-0.08	1.05	1269	1.00
Augmented:BlockOrder	0.57	0.32	-0.04	1.17	1589	1.00

Significant population-level effects for Model 3.2 but with ScaleTone instead of SPCS along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence,  $Rhat = 1$ ). The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

## C.2.2 Model 3.3

TABLE C.8: All population-level effects for Model 3.3, which was summarized in Table 3.9

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.40	0.11	-3.61	-3.19	5108	1
Intercept[2]	-2.07	0.10	-2.28	-1.86	4793	1
Intercept[3]	-0.89	0.10	-1.10	-0.69	4533	1
Intercept[4]	-0.19	0.10	-0.39	0.01	4478	1
Intercept[5]	1.00	0.10	0.80	1.20	4492	1
Intercept[6]	2.48	0.11	2.27	2.68	4739	1
MusSoph	0.20	0.09	0.02	0.37	4660	1

*Continued on next page*

Table C.8 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Height	-0.08	0.18	-0.42	0.28	7887	1
RelHeight	-0.01	0.18	-0.35	0.34	8334	1
Height <sup>2</sup>	0.03	0.08	-0.13	0.19	8942	1
RelHeight <sup>2</sup>	-0.02	0.08	-0.18	0.14	9118	1
Primacy	-0.01	0.05	-0.10	0.09	16728	1
Previous	0.29	0.04	0.21	0.37	8977	1
Recency	0.09	0.06	-0.03	0.20	12412	1
SPCS	0.37	0.04	0.28	0.45	8689	1
MelCont	0.04	0.02	-0.01	0.09	15848	1
Count	-0.01	0.05	-0.11	0.10	10996	1
Minor	-0.70	0.09	-0.87	-0.52	6551	1
Diminished	-1.10	0.11	-1.32	-0.87	6302	1
Augmented	-1.43	0.14	-1.72	-1.14	8986	1
TrialNo	0.08	0.05	-0.03	0.19	10166	1
InContTrialNo	-0.09	0.04	-0.17	-0.01	11907	1
TrialNo <sup>2</sup>	0.06	0.04	-0.02	0.15	11128	1
BlockOrder	-0.26	0.19	-0.64	0.12	4230	1
MusSoph:Height	0.02	0.16	-0.28	0.33	9609	1
MusSoph:RelHeight	0.06	0.15	-0.23	0.35	10521	1
MusSoph:Height <sup>2</sup>	0.03	0.07	-0.10	0.16	11056	1
MusSoph:RelHeight <sup>2</sup>	-0.03	0.07	-0.16	0.10	10618	1
MusSoph:Primacy	-0.03	0.04	-0.10	0.04	19802	1
MusSoph:Previous	-0.03	0.04	-0.10	0.04	7792	1
MusSoph:Recency	0.02	0.05	-0.07	0.11	12156	1
MusSoph:SPCS	0.22	0.04	0.15	0.30	8247	1
MusSoph:MelCont	0.01	0.02	-0.02	0.05	19794	1
MusSoph:Count	0.08	0.03	0.01	0.15	12028	1
MusSoph:Minor	-0.13	0.08	-0.30	0.03	6080	1
MusSoph:Diminished	-0.08	0.11	-0.30	0.13	5888	1
MusSoph:Augmented	-0.40	0.12	-0.64	-0.16	6965	1
Height:TrialNo	0.35	0.17	0.01	0.68	8265	1
RelHeight:TrialNo	-0.36	0.17	-0.70	-0.03	8289	1
Height <sup>2</sup> :TrialNo	0.01	0.08	-0.14	0.16	10049	1
RelHeight <sup>2</sup> :TrialNo	-0.02	0.08	-0.17	0.13	10088	1
Primacy:TrialNo	0.01	0.04	-0.07	0.09	17726	1
Previous:TrialNo	-0.03	0.02	-0.07	0.02	15427	1
Recency:TrialNo	0.01	0.04	-0.07	0.09	14872	1
SPCS:TrialNo	0.02	0.03	-0.03	0.07	12904	1
MelCont:TrialNo	0.00	0.02	-0.04	0.04	15116	1
Count:TrialNo	-0.01	0.04	-0.08	0.06	11465	1
Minor:TrialNo	-0.09	0.05	-0.19	0.00	11776	1
Diminished:TrialNo	-0.12	0.05	-0.23	-0.01	13512	1
Augmented:TrialNo	-0.03	0.08	-0.20	0.14	13148	1
Height:InContTrialNo	-0.08	0.11	-0.29	0.13	10150	1
RelHeight:InContTrialNo	0.05	0.11	-0.16	0.27	10173	1
Height <sup>2</sup> :InContTrialNo	-0.01	0.05	-0.11	0.10	11330	1
RelHeight <sup>2</sup> :InContTrialNo	0.01	0.05	-0.10	0.12	11338	1
Primacy:InContTrialNo	0.06	0.04	-0.01	0.13	16935	1
Previous:InContTrialNo	0.01	0.02	-0.03	0.06	13662	1
Recency:InContTrialNo	-0.02	0.04	-0.09	0.05	16655	1
SPCS:InContTrialNo	-0.02	0.02	-0.06	0.02	20102	1
MelCont:InContTrialNo	0.02	0.02	-0.02	0.05	17086	1

Continued on next page



Table C.8 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Count:lnContTrialNo	-0.05	0.02	-0.09	-0.01	17666	1
Minor:lnContTrialNo	0.09	0.04	0.01	0.17	14081	1
Diminished:lnContTrialNo	0.12	0.05	0.03	0.21	13882	1
Augmented:lnContTrialNo	0.09	0.07	-0.04	0.22	16580	1
Height:TrialNo <sup>2</sup>	0.20	0.14	-0.07	0.47	7257	1
RelHeight:TrialNo <sup>2</sup>	-0.21	0.14	-0.48	0.06	7211	1
Height <sup>2</sup> :TrialNo <sup>2</sup>	-0.01	0.06	-0.13	0.11	8371	1
RelHeight <sup>2</sup> :TrialNo <sup>2</sup>	0.00	0.06	-0.12	0.12	8416	1
Primacy:TrialNo <sup>2</sup>	-0.02	0.04	-0.10	0.05	13753	1
Previous:TrialNo <sup>2</sup>	0.01	0.02	-0.03	0.05	16402	1
Recency:TrialNo <sup>2</sup>	0.05	0.04	-0.02	0.13	14105	1
SPCS:TrialNo <sup>2</sup>	0.03	0.02	-0.01	0.08	10876	1
MelCont:TrialNo <sup>2</sup>	0.01	0.02	-0.03	0.04	13179	1
Count:TrialNo <sup>2</sup>	-0.03	0.02	-0.08	0.01	10469	1
Minor:TrialNo <sup>2</sup>	0.01	0.04	-0.07	0.09	13544	1
Diminished:TrialNo <sup>2</sup>	-0.01	0.05	-0.12	0.09	12180	1
Augmented:TrialNo <sup>2</sup>	-0.03	0.08	-0.19	0.12	14743	1
Height:BlockOrder	-0.11	0.30	-0.70	0.48	7896	1
RelHeight:BlockOrder	0.19	0.29	-0.39	0.76	8407	1
Height <sup>2</sup> :BlockOrder	-0.06	0.13	-0.32	0.19	10437	1
RelHeight <sup>2</sup> :BlockOrder	0.10	0.13	-0.16	0.35	9929	1
Primacy:BlockOrder	0.01	0.07	-0.13	0.15	18781	1
Previous:BlockOrder	-0.01	0.07	-0.16	0.14	7376	1
Recency:BlockOrder	-0.02	0.10	-0.22	0.17	12045	1
SPCS:BlockOrder	-0.04	0.08	-0.20	0.12	7784	1
MelCont:BlockOrder	0.03	0.03	-0.03	0.10	21670	1
Count:BlockOrder	-0.27	0.09	-0.44	-0.10	10861	1
Minor:BlockOrder	0.20	0.17	-0.13	0.52	6322	1
Diminished:BlockOrder	0.23	0.22	-0.20	0.65	5738	1
Augmented:BlockOrder	0.14	0.25	-0.34	0.63	7282	1

TABLE C.9: Model 3.3 but with ScaleTone instead of SPCS significant population-level effects

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-3.37	0.10	-3.57	-3.17	535	1.01
Intercept[2]	-2.04	0.10	-2.24	-1.84	522	1.01
Intercept[3]	-0.87	0.10	-1.06	-0.67	521	1.01
Intercept[4]	-0.16	0.10	-0.36	0.03	512	1.01
Intercept[5]	1.02	0.10	0.82	1.21	535	1.01
Intercept[6]	2.49	0.10	2.29	2.69	582	1.01
MusSoph	0.21	0.09	0.02	0.39	352	1.01
Previous	0.29	0.04	0.21	0.37	559	1.00
ScaleTone	0.32	0.04	0.23	0.41	595	1.00
Count	0.01	0.06	-0.09	0.12	1107	1.00
Minor	-0.69	0.09	-0.86	-0.53	448	1.01
Diminished	-1.10	0.11	-1.33	-0.87	333	1.01
Augmented	-1.43	0.14	-1.70	-1.16	495	1.01
TrialNo	0.09	0.06	-0.02	0.20	977	1.00
lnContTrialNo	-0.10	0.04	-0.18	-0.02	930	1.00
BlockOrder	-0.22	0.20	-0.61	0.17	436	1.00
MusSoph:ScaleTone	0.21	0.04	0.13	0.29	674	1.01
MusSoph:Count	0.09	0.04	0.02	0.16	1025	1.00
MusSoph:Minor	-0.13	0.08	-0.29	0.05	440	1.00
MusSoph:Diminished	-0.09	0.11	-0.30	0.13	535	1.00
MusSoph:Augmented	-0.39	0.12	-0.64	-0.15	476	1.00
Minor:TrialNo	-0.11	0.05	-0.20	-0.01	1000	1.00
Diminished:TrialNo	-0.11	0.05	-0.21	-0.01	1314	1.00
Augmented:TrialNo	-0.07	0.08	-0.23	0.09	1386	1.00
Count:lnContTrialNo	-0.05	0.02	-0.09	-0.01	1392	1.00
Minor:lnContTrialNo	0.10	0.04	0.03	0.18	1288	1.00
Diminished:lnContTrialNo	0.14	0.04	0.05	0.22	1568	1.00
Augmented:lnContTrialNo	0.11	0.07	-0.03	0.24	1520	1.00
Count:Order	-0.30	0.09	-0.49	-0.14	852	1.00

Significant population-level effects for Model 3.3 but with ScaleTone instead of SPCS along with intercepts and conditional main effects for significant interaction effects. *CI* stands for Credibility Interval. *Estimate* refers to the coefficient for the associated effect in the model. Interactions between effects are indicated with ':' written between the interacting effects. As written in *brms*, 'for each parameter, *Eff. Sample* is a crude measure of effective sample size, and *Rhat* is the potential scale reduction factor on split chains (at convergence, *Rhat* = 1)'. The chains are obtained via *Markov chain Monte Carlo (MCMC)*, which are a class of algorithms used in *brms* for sampling from a probability distribution.

## Appendix D

# Experiments 4 & 5

## D.1 Experiment 4 Confirmatory

TABLE D.1: All population-level effects for Model Tone, which was summarized in Table 6.3

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept <sup>[1]</sup>	-2.86	0.09	-3.03	-2.69	5373	1
Intercept <sup>[2]</sup>	-1.57	0.08	-1.74	-1.41	5130	1
Intercept <sup>[3]</sup>	-0.41	0.08	-0.57	-0.24	4995	1
Intercept <sup>[4]</sup>	0.24	0.08	0.07	0.40	4998	1
Intercept <sup>[5]</sup>	1.27	0.08	1.11	1.44	5104	1
Intercept <sup>[6]</sup>	2.61	0.09	2.43	2.78	5428	1
MusSoph	-0.13	0.08	-0.28	0.02	5453	1
Height	0.24	0.17	-0.08	0.56	10016	1
RelHeight	-0.14	0.16	-0.46	0.18	10510	1
Height <sup>2</sup>	-0.08	0.09	-0.25	0.09	9613	1
RelHeight <sup>2</sup>	-0.03	0.09	-0.2	0.14	9779	1
Primacy	-0.05	0.09	-0.23	0.14	19504	1
Previous	0.25	0.05	0.14	0.36	9006	1
Recency	0.30	0.12	0.06	0.54	14460	1
SPCS	0.46	0.06	0.34	0.59	9169	1
MelCont	-0.02	0.03	-0.07	0.02	23087	1
Count	0.01	0.04	-0.07	0.09	16965	1
TrialNo	0.03	0.04	-0.04	0.1	16257	1
lnContTrialNo	-0.02	0.02	-0.07	0.02	22471	1
TrialNo <sup>2</sup>	-0.04	0.04	-0.12	0.05	15177	1
MusSoph:Height	0.13	0.11	-0.08	0.35	13323	1
MusSoph:RelHeight	-0.13	0.11	-0.34	0.07	14206	1
MusSoph:Height <sup>2</sup>	0.01	0.05	-0.1	0.11	12795	1
MusSoph:RelHeight <sup>2</sup>	0.04	0.05	-0.06	0.14	12438	1
MusSoph:Primacy	-0.01	0.07	-0.14	0.12	20616	1
MusSoph:Previous	0.00	0.05	-0.09	0.09	9485	1
MusSoph:Recency	0.37	0.10	0.17	0.58	13741	1
MusSoph:SPCS	0.29	0.06	0.18	0.40	9957	1
MusSoph:MelCont	-0.01	0.02	-0.04	0.02	25526	1
MusSoph:Count	0.04	0.03	-0.02	0.10	13952	1
Height:TrialNo	-0.2	0.10	-0.39	0.00	10259	1
RelHeight:TrialNo	0.14	0.10	-0.06	0.33	10179	1

*Continued on next page*

Table D.1 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Height <sup>2</sup> :TrialNo	0.03	0.05	-0.07	0.13	17238	1
RelHeight <sup>2</sup> :TrialNo	-0.03	0.05	-0.13	0.07	17781	1
Primacy: TrialNo	0.10	0.07	-0.03	0.23	23839	1
Previous: TrialNo	-0.03	0.02	-0.07	0.01	22147	1
Recency: TrialNo	-0.06	0.07	-0.21	0.08	20903	1
SPCS: TrialNo	-0.04	0.02	-0.09	0.00	16660	1
MelCont: TrialNo	0.00	0.02	-0.04	0.04	18714	1
Count: TrialNo	0.08	0.04	-0.01	0.16	17706	1
Height:lnContTrialNo	-0.03	0.08	-0.19	0.12	19576	1
RelHeight:lnContTrialNo	0.03	0.08	-0.12	0.18	19298	1
Height <sup>2</sup> :lnContTrialNo	0.03	0.04	-0.06	0.11	18488	1
RelHeight <sup>2</sup> :lnContTrialNo	-0.04	0.04	-0.13	0.04	18656	1
Primacy:lnContTrialNo	0.01	0.06	-0.12	0.13	23275	1
Previous:lnContTrialNo	0.02	0.02	-0.03	0.06	15918	1
Recency:lnContTrialNo	0.01	0.06	-0.12	0.13	21462	1
SPCS:lnContTrialNo	0.04	0.02	0.00	0.08	17731	1
MelCont:lnContTrialNo	0.02	0.02	-0.01	0.06	22158	1
Count:lnContTrialNo	0.02	0.02	-0.02	0.06	19165	1
Height: TrialNo <sup>2</sup>	-0.21	0.13	-0.46	0.04	9826	1
RelHeight: TrialNo <sup>2</sup>	0.22	0.13	-0.03	0.46	9812	1
Height <sup>2</sup> : TrialNo <sup>2</sup>	-0.02	0.06	-0.14	0.11	9446	1
RelHeight <sup>2</sup> : TrialNo <sup>2</sup>	0.00	0.06	-0.13	0.13	9255	1
Primacy: TrialNo <sup>2</sup>	-0.04	0.07	-0.19	0.10	17536	1
Previous: TrialNo <sup>2</sup>	0.00	0.03	-0.05	0.06	13859	1
Recency: TrialNo <sup>2</sup>	0.12	0.09	-0.06	0.29	15788	1
SPCS: TrialNo <sup>2</sup>	0.01	0.03	-0.04	0.06	14285	1
MelCont: TrialNo <sup>2</sup>	0.02	0.02	-0.02	0.05	21047	1
Count: TrialNo <sup>2</sup>	-0.04	0.03	-0.1	0.01	19074	1
RelHeight <sup>2</sup> : TrialNo <sup>2</sup>	0.00	0.08	-0.17	0.16	6501	1
Primacy: TrialNo <sup>2</sup>	-0.06	0.08	-0.22	0.09	12450	1
Previous: TrialNo <sup>2</sup>	0.01	0.03	-0.04	0.06	11724	1
Recency: TrialNo <sup>2</sup>	0.14	0.09	-0.05	0.32	10519	1
SPCS: TrialNo <sup>2</sup>	0.01	0.03	-0.04	0.06	9768	1
MelCont: TrialNo <sup>2</sup>	0.02	0.02	-0.02	0.06	15961	1
Count: TrialNo <sup>2</sup>	-0.03	0.04	-0.1	0.04	7084	1

## D.2 Experiment 5 Confirmatory

TABLE D.2: All population-level effects for Model Triad, which was summarized in Table 6.4

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.54	0.11	-2.76	-2.32	1567	1
Intercept[2]	-1.22	0.11	-1.43	-1	1525	1
Intercept[3]	-0.18	0.11	-0.38	0.04	1513	1
Intercept[4]	0.47	0.11	0.26	0.69	1517	1

Continued on next page

Table D.2 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept <sup>[5]</sup>	1.62	0.11	1.4	1.83	1536	1
Intercept <sup>[6]</sup>	3.05	0.11	2.83	3.28	1624	1
MusSoph	-0.04	0.10	-0.24	0.15	1503	1
Height	-0.12	0.14	-0.39	0.16	3648	1
RelHeight	0.00	0.13	-0.25	0.25	4063	1
Height <sup>2</sup>	-0.03	0.05	-0.14	0.07	4233	1
RelHeight <sup>2</sup>	0.00	0.05	-0.1	0.10	4106	1
Primacy	-0.02	0.06	-0.14	0.11	7680	1
Previous	0.36	0.07	0.24	0.49	2863	1
Recency	0.05	0.06	-0.08	0.18	7579	1
SPCS	0.28	0.05	0.18	0.39	3719	1
MelCont	0.06	0.03	0.01	0.11	8407	1
Count	-0.01	0.05	-0.11	0.08	7082	1
ChordStab	0.64	0.08	0.48	0.80	3001	1
TrialNo	0.05	0.06	-0.06	0.16	4450	1
InContTrialNo	0.07	0.03	0.01	0.13	8247	1
TrialNo <sup>2</sup>	0.00	0.06	-0.11	0.10	4825	1
MusSoph:Height	0.07	0.11	-0.16	0.30	3237	1
MusSoph:RelHeight	-0.08	0.10	-0.27	0.12	4023	1
MusSoph:Height <sup>2</sup>	-0.03	0.04	-0.1	0.04	6925	1
MusSoph:RelHeight <sup>2</sup>	0.01	0.04	-0.07	0.08	6173	1
MusSoph:Primacy	-0.04	0.04	-0.13	0.05	12486	1
MusSoph:Previous	-0.01	0.06	-0.13	0.11	3092	1
MusSoph:Recency	0.02	0.05	-0.08	0.11	9381	1
MusSoph:SPCS	0.17	0.05	0.09	0.26	4029	1
MusSoph:MelCont	0.01	0.02	-0.03	0.04	14085	1
MusSoph:Count	-0.02	0.03	-0.07	0.05	6501	1
MusSoph:ChordStab	0.14	0.08	-0.01	0.29	3363	1
Height:TrialNo	0.05	0.08	-0.11	0.21	4971	1
RelHeight:TrialNo	-0.06	0.08	-0.22	0.09	4913	1
Height <sup>2</sup> :TrialNo	-0.06	0.04	-0.13	0.01	6230	1
RelHeight <sup>2</sup> :TrialNo	0.03	0.03	-0.04	0.09	6322	1
Primacy:TrialNo	-0.02	0.04	-0.1	0.07	12536	1
Previous:TrialNo	-0.01	0.03	-0.07	0.04	6895	1
Recency:TrialNo	0.04	0.05	-0.05	0.14	9701	1
SPCS:TrialNo	-0.08	0.02	-0.13	-0.03	8192	1
MelCont:TrialNo	0.01	0.02	-0.03	0.04	11451	1
Count:TrialNo	-0.03	0.05	-0.12	0.07	4531	1
ChordStab:TrialNo	0.00	0.03	-0.06	0.05	7486	1
Height:InContTrialNo	-0.04	0.06	-0.15	0.07	6572	1
RelHeight:InContTrialNo	0.05	0.06	-0.06	0.16	6577	1
Height <sup>2</sup> :InContTrialNo	-0.02	0.03	-0.08	0.04	6992	1
RelHeight <sup>2</sup> :InContTrialNo	0.00	0.03	-0.06	0.06	7239	1
Primacy:InContTrialNo	0.03	0.04	-0.06	0.11	10368	1
Previous:InContTrialNo	0.00	0.02	-0.04	0.04	11828	1
Recency:InContTrialNo	-0.11	0.04	-0.19	-0.02	9214	1
SPCS:InContTrialNo	0.05	0.02	0.01	0.09	7797	1
MelCont:InContTrialNo	0.00	0.02	-0.04	0.03	12333	1
Count:InContTrialNo	-0.04	0.02	-0.08	0.00	10775	1
ChordStab:InContTrialNo	-0.05	0.02	-0.09	-0.01	9586	1
Height:TrialNo <sup>2</sup>	0.06	0.10	-0.13	0.25	3915	1
RelHeight:TrialNo <sup>2</sup>	-0.01	0.10	-0.2	0.18	3824	1

Continued on next page

Table D.2 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Height <sup>2</sup> :TrialNo <sup>2</sup>	0.02	0.04	-0.06	0.10	3819	1
RelHeight <sup>2</sup> :TrialNo <sup>2</sup>	0.01	0.04	-0.07	0.08	3887	1
Primacy: TrialNo <sup>2</sup>	-0.01	0.05	-0.1	0.08	7513	1
Previous: TrialNo <sup>2</sup>	0.02	0.03	-0.04	0.07	8665	1
Recency: TrialNo <sup>2</sup>	0.02	0.05	-0.08	0.13	6897	1
SPCS: TrialNo <sup>2</sup>	0.02	0.03	-0.03	0.07	6915	1
MelCont: TrialNo <sup>2</sup>	-0.02	0.02	-0.05	0.02	8318	1
Count: TrialNo <sup>2</sup>	0.01	0.03	-0.05	0.06	7120	1
ChordStab: TrialNo <sup>2</sup>	-0.01	0.02	-0.05	0.04	9927	1

### D.3 Experiment 4 Exploratory 1

TABLE D.3: All population-level effects for Exploratory Model 6.1, which was summarized in Table 6.6

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept <sup>[1]</sup>	-3.01	0.12	-3.24	-2.78	3467	1
Intercept <sup>[2]</sup>	-1.65	0.11	-1.87	-1.43	3279	1
Intercept <sup>[3]</sup>	-0.42	0.11	-0.63	-0.2	3257	1
Intercept <sup>[4]</sup>	0.26	0.11	0.05	0.48	3277	1
Intercept <sup>[5]</sup>	1.35	0.11	1.14	1.57	3297	1
Intercept <sup>[6]</sup>	2.75	0.11	2.53	2.97	3463	1
MusSoph	-0.11	0.08	-0.28	0.05	2900	1
Height	0.26	0.57	-0.88	1.36	1957	1
RelHeight	-0.09	0.57	-1.2	1.05	1964	1
Height <sup>2</sup>	0.38	0.44	-0.47	1.28	1413	1
RelHeight <sup>2</sup>	-0.49	0.45	-1.4	0.38	1406	1
Primacy	-0.01	0.18	-0.36	0.34	5202	1
Previous	0.17	0.07	0.03	0.31	4843	1
Recency	0.38	0.20	-0.01	0.76	5101	1
SPCS	0.46	0.08	0.31	0.62	3554	1
MelCont	-0.01	0.05	-0.11	0.09	4674	1
Count	0.03	0.07	-0.1	0.16	5102	1
TrialNo	0.03	0.04	-0.05	0.11	9627	1
InContTrialNo	-0.03	0.02	-0.08	0.02	18079	1
TrialNo <sup>2</sup>	-0.01	0.04	-0.09	0.07	9127	1
Scale2	-0.1	0.11	-0.31	0.11	7890	1
Scale3	0.13	0.10	-0.07	0.33	7496	1
Scale4	-0.06	0.13	-0.32	0.19	6990	1
Scale5	0.08	0.12	-0.16	0.32	7459	1
Scale6	-0.01	0.10	-0.2	0.18	7309	1
Scale7	-0.07	0.12	-0.31	0.16	6994	1
Scale8	0.01	0.10	-0.19	0.20	7746	1
MusSoph:Height	0.12	0.13	-0.14	0.38	6835	1
MusSoph:RelHeight	-0.13	0.13	-0.38	0.11	7564	1
MusSoph:Height <sup>2</sup>	0.00	0.07	-0.13	0.13	6314	1
MusSoph:RelHeight <sup>2</sup>	0.06	0.07	-0.07	0.19	6438	1

Continued on next page

Table D.3 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
MusSoph:Primacy	-0.02	0.07	-0.16	0.12	13706	1
MusSoph:Previous	0.02	0.05	-0.08	0.11	4513	1
MusSoph:Recency	0.40	0.11	0.18	0.62	8644	1
MusSoph:SPCS	0.29	0.06	0.17	0.41	3986	1
MusSoph:MelCont	-0.01	0.02	-0.05	0.03	16744	1
MusSoph:Count	0.04	0.03	-0.02	0.10	8366	1
Height: TrialNo	-0.01	0.13	-0.25	0.24	4357	1
RelHeight: TrialNo	-0.05	0.13	-0.3	0.19	4400	1
Height <sup>2</sup> : TrialNo	-0.01	0.06	-0.14	0.11	7681	1
RelHeight <sup>2</sup> : TrialNo	0.01	0.06	-0.12	0.13	7258	1
Primacy: TrialNo	0.12	0.07	-0.02	0.26	14862	1
Previous: TrialNo	-0.03	0.02	-0.08	0.01	13804	1
Recency: TrialNo	-0.12	0.08	-0.28	0.04	13904	1
SPCS: TrialNo	-0.04	0.03	-0.09	0.02	8148	1
MelCont: TrialNo	0.00	0.02	-0.04	0.04	16279	1
Count: TrialNo	0.05	0.05	-0.05	0.15	7661	1
Height: lnCont TrialNo	-0.06	0.08	-0.22	0.10	10455	1
RelHeight: lnCont TrialNo	0.05	0.08	-0.11	0.21	10452	1
Height <sup>2</sup> : lnCont TrialNo	0.04	0.05	-0.05	0.13	9781	1
RelHeight <sup>2</sup> : lnCont TrialNo	-0.06	0.05	-0.15	0.03	10255	1
Primacy: lnCont TrialNo	0.01	0.07	-0.12	0.14	15640	1
Previous: lnCont TrialNo	0.01	0.02	-0.03	0.06	12057	1
Recency: lnCont TrialNo	0.06	0.07	-0.07	0.19	17568	1
SPCS: lnCont TrialNo	0.04	0.02	-0.01	0.08	13875	1
MelCont: lnCont TrialNo	0.02	0.02	-0.01	0.06	17053	1
Count: lnCont TrialNo	0.02	0.02	-0.02	0.07	13131	1
Height: TrialNo <sup>2</sup>	-0.22	0.16	-0.53	0.10	6282	1
RelHeight: TrialNo <sup>2</sup>	0.23	0.16	-0.09	0.55	6180	1
lHeight <sup>2</sup> : TrialNo <sup>2</sup>	-0.03	0.08	-0.19	0.14	6541	1
lRelHeight <sup>2</sup> : TrialNo <sup>2</sup>	0.00	0.08	-0.17	0.16	6501	1
Primacy: TrialNo <sup>2</sup>	-0.06	0.08	-0.22	0.09	12450	1
Previous: TrialNo <sup>2</sup>	0.01	0.03	-0.04	0.06	11724	1
Recency: TrialNo <sup>2</sup>	0.14	0.09	-0.05	0.32	10519	1
SPCS: TrialNo <sup>2</sup>	0.01	0.03	-0.04	0.06	9768	1
MelCont: TrialNo <sup>2</sup>	0.02	0.02	-0.02	0.06	15961	1
Count: TrialNo <sup>2</sup>	-0.03	0.04	-0.1	0.04	7084	1
Height: Scale2	0.36	0.85	-1.28	2.03	3872	1
Height: Scale3	0.22	0.81	-1.33	1.79	3689	1
Height: Scale4	0.42	0.83	-1.19	2.06	3775	1
Height: Scale5	0.60	0.70	-0.77	1.98	2578	1
Height: Scale6	0.13	0.60	-1.03	1.32	2124	1
Height: Scale7	-0.11	0.63	-1.32	1.13	2317	1
Height: Scale8	-0.29	0.56	-1.38	0.82	1901	1
RelHeight: Scale2	-0.49	0.86	-2.18	1.16	3865	1
RelHeight: Scale3	-0.16	0.81	-1.75	1.42	3731	1
RelHeight: Scale4	-0.56	0.84	-2.21	1.06	3766	1
RelHeight: Scale5	-0.72	0.70	-2.11	0.65	2579	1
RelHeight: Scale6	-0.18	0.60	-1.38	0.99	2151	1
RelHeight: Scale7	0.02	0.63	-1.23	1.24	2306	1
RelHeight: Scale8	0.14	0.57	-0.99	1.23	1897	1
Height <sup>2</sup> : Scale2	0.00	0.55	-1.09	1.07	2020	1

Continued on next page

Table D.3 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Height <sup>2</sup> :Scale3	-0.32	0.54	-1.4	0.73	1824	1
Height <sup>2</sup> :Scale4	-0.67	0.52	-1.69	0.32	1848	1
Height <sup>2</sup> :Scale5	-0.43	0.48	-1.39	0.49	1645	1
Height <sup>2</sup> :Scale6	-0.6	0.47	-1.52	0.31	1443	1
Height <sup>2</sup> :Scale7	-0.34	0.46	-1.26	0.55	1574	1
Height <sup>2</sup> :Scale8	-0.41	0.44	-1.31	0.45	1372	1
RelHeight <sup>2</sup> :Scale2	-0.02	0.56	-1.12	1.07	2018	1
RelHeight <sup>2</sup> :Scale3	0.24	0.54	-0.83	1.33	1840	1
RelHeight <sup>2</sup> :Scale4	0.68	0.53	-0.32	1.72	1879	1
RelHeight <sup>2</sup> :Scale5	0.47	0.48	-0.46	1.44	1641	1
RelHeight <sup>2</sup> :Scale6	0.59	0.47	-0.32	1.53	1443	1
RelHeight <sup>2</sup> :Scale7	0.40	0.46	-0.5	1.33	1551	1
RelHeight <sup>2</sup> :Scale8	0.44	0.44	-0.42	1.34	1379	1
Primacy:Scale2	-0.24	0.25	-0.73	0.24	6599	1
Primacy:Scale3	0.27	0.26	-0.25	0.79	6879	1
Primacy:Scale4	0.10	0.25	-0.38	0.60	6767	1
Primacy:Scale5	-0.01	0.25	-0.49	0.47	6646	1
Primacy:Scale6	-0.07	0.26	-0.58	0.44	7290	1
Primacy:Scale7	-0.18	0.26	-0.69	0.33	7146	1
Primacy:Scale8	-0.01	0.26	-0.53	0.50	6811	1
Previous:Scale2	0.16	0.08	0.01	0.32	6842	1
Previous:Scale3	-0.01	0.10	-0.19	0.19	7468	1
Previous:Scale4	-0.08	0.08	-0.23	0.07	7388	1
Previous:Scale5	0.12	0.08	-0.03	0.28	6853	1
Previous:Scale6	0.02	0.07	-0.13	0.16	7603	1
Previous:Scale7	0.11	0.09	-0.05	0.28	6454	1
Previous:Scale8	0.03	0.08	-0.12	0.19	7978	1
Recency:Scale2	-0.53	0.27	-1.05	0.00	7246	1
Recency:Scale3	0.12	0.27	-0.4	0.65	6816	1
Recency:Scale4	-0.11	0.27	-0.62	0.42	6591	1
Recency:Scale5	0.43	0.30	-0.15	1.04	7062	1
Recency:Scale6	-0.22	0.26	-0.74	0.28	6664	1
Recency:Scale7	-0.07	0.28	-0.62	0.47	7022	1
Recency:Scale8	-0.04	0.28	-0.59	0.51	6431	1
SPCS:Scale2	0.30	0.08	0.14	0.46	5976	1
SPCS:Scale3	-0.01	0.08	-0.17	0.15	6116	1
SPCS:Scale4	0.07	0.08	-0.09	0.23	5967	1
SPCS:Scale5	-0.14	0.09	-0.32	0.05	6672	1
SPCS:Scale6	-0.09	0.08	-0.24	0.06	6085	1
SPCS:Scale7	-0.03	0.09	-0.2	0.14	6760	1
SPCS:Scale8	0.11	0.09	-0.06	0.30	6178	1
MelCont:Scale2	-0.04	0.07	-0.18	0.09	5861	1
MelCont:Scale3	-0.09	0.07	-0.24	0.05	6299	1
MelCont:Scale4	-0.06	0.07	-0.19	0.08	6193	1
MelCont:Scale5	0.04	0.07	-0.1	0.17	6479	1
MelCont:Scale6	0.01	0.07	-0.12	0.15	6127	1
MelCont:Scale7	-0.03	0.07	-0.17	0.11	6492	1
MelCont:Scale8	0.03	0.08	-0.12	0.17	6301	1
Count:Scale2	-0.18	0.10	-0.38	0.02	6478	1
Count:Scale3	-0.14	0.09	-0.31	0.03	6588	1
Count:Scale4	0.15	0.08	-0.02	0.31	6522	1
Count:Scale5	0.02	0.08	-0.13	0.18	6623	1

Continued on next page



Table D.3 – *Continued from previous page*

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Count:Scale6	-0.08	0.10	-0.27	0.12	6243	1
Count:Scale7	0.07	0.08	-0.1	0.23	6145	1
Count:Scale8	-0.07	0.08	-0.23	0.09	7021	1

## D.4 Experiment 4 Exploratory 2

TABLE D.4: All population-level effects for Exploratory Model 6.2

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.58	1.35	-5.23	0.20	5514	1
Intercept[2]	-1.37	1.35	-4.02	1.41	5507	1
Intercept[3]	-0.27	1.35	-2.91	2.52	5513	1
Intercept[4]	0.34	1.35	-2.3	3.13	5520	1
Intercept[5]	1.33	1.35	-1.3	4.11	5515	1
Intercept[6]	2.62	1.35	-0.02	5.4	5529	1
MusSoph	-0.06	0.06	-0.18	0.07	1127	1
SGC	-0.05	2.56	-5.42	5.14	4989	1
TriadEnt	0.11	2	-3.83	4.14	4016	1
MaxVar	-0.04	1.6	-3.19	3.16	3121	1
WF	0.08	3.14	-6.1	6.76	6901	1
PWF	-0.09	2.04	-4.16	4	3705	1
TrichordEnt	-0.01	1.59	-3.16	3.13	3658	1
PentachordEnt	0.03	2.32	-4.66	4.68	5053	1
LummaStblty	0.05	1.49	-2.82	3.05	2361	1
Lummalmppty	0.02	1.25	-2.5	2.46	2851	1
MaxCons	0.01	0.93	-1.8	1.89	2331	1
MinCons	-0.01	1.75	-3.53	3.36	3411	1
MedCons	-0.07	1.91	-3.84	3.78	2750	1
Tetrachrdlty	-0.01	0.99	-1.92	2.01	2263	1
Previous	0.18	1.15	-2.12	2.43	4135	1
Recency	0.35	1.11	-1.89	2.51	3872	1
SPCS	0.34	1.13	-1.91	2.62	3862	1
MusSoph:Previous	0.02	0.04	-0.07	0.11	2948	1
MusSoph:Recency	0.32	0.11	0.11	0.53	3978	1
MusSoph:SPCS	0.25	0.05	0.15	0.35	3237	1
SGC:Previous	0.03	2.49	-4.95	5.2	5775	1
SGC:Recency	-0.08	2.57	-5.27	5.14	5514	1
SGC:SPCS	0.11	2.53	-4.96	5.19	4674	1
TriadEnt:Previous	-0.04	1.93	-3.79	3.8	4407	1
TriadEnt:Recency	0.11	1.93	-3.68	3.99	4305	1
TriadEnt:SPCS	0.01	1.91	-3.79	3.79	3940	1
MaxVar:Previous	0.00	1.57	-3.05	3.13	3002	1
MaxVar:Recency	0.01	1.57	-3.08	3.19	3468	1
MaxVar:SPCS	-0.07	1.54	-3.11	2.97	3297	1
WF:Previous	0.10	2.66	-5.16	5.57	5269	1
WF:Recency	0.28	2.61	-4.91	5.54	4015	1
WF:SPCS	0.21	2.64	-4.95	5.54	4179	1
PWF:Previous	0.07	1.96	-3.75	4.13	3186	1
PWF:Recency	-0.1	1.9	-3.85	3.65	4032	1

*Continued on next page*

Table D.4 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
PWF:SPCS	0.16	1.9	-3.6	3.93	3733	1
TrichordEnt:Previous	0.08	1.56	-3.02	3.16	3877	1
TrichordEnt:Recency	-0.03	1.6	-3.24	3.14	3834	1
TrichordEnt:SPCS	0.04	1.58	-3.19	3.13	3535	1
PentachordEnt:Previous	-0.01	2.25	-4.59	4.5	5928	1
PentachordEnt:Recency	0.01	2.23	-4.5	4.44	5510	1
PentachordEnt:SPCS	0.01	2.25	-4.52	4.54	5572	1
LummaStblty:Previous	-0.05	1.49	-3.06	2.96	3082	1
LummaStblty:Recency	0.06	1.45	-2.81	2.96	2315	1
LummaStblty:SPCS	-0.01	1.48	-2.98	2.92	2576	1
LummaImprty:Previous	0.05	1.23	-2.4	2.45	2843	1
LummaImprty:Recency	-0.19	1.22	-2.6	2.2	3122	1
LummaImprty:SPCS	-0.08	1.19	-2.39	2.28	2911	1
MaxCons:Previous	0.09	0.92	-1.76	1.91	2855	1
MaxCons:Recency	-0.04	0.92	-1.85	1.77	2288	1
MaxCons:SPCS	-0.07	0.90	-1.91	1.69	2334	1
MinCons:Previous	-0.04	1.74	-3.48	3.46	3708	1
MinCons:Recency	-0.1	1.78	-3.69	3.34	3171	1
MinCons:SPCS	0.01	1.75	-3.43	3.51	3177	1
MedCons:Previous	0.07	1.88	-3.63	3.79	3988	1
MedCons:Recency	-0.17	1.85	-3.81	3.55	3499	1
MedCons:SPCS	0.04	1.83	-3.58	3.7	3386	1
Tetrachrdlty:Previous	-0.04	1	-2.04	1.93	2814	1
Tetrachrdlty:Recency	-0.15	0.98	-2.06	1.77	2164	1
Tetrachrdlty:SPCS	0.02	0.97	-1.96	1.96	2264	1

## D.5 Experiment 4 Exploratory 3

TABLE D.5: All population-level effects for Exploratory Model 6.1, which was summarized in Table 6.8

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Intercept[1]	-2.92	0.09	-3.1	-2.74	3069	1
Intercept[2]	-1.61	0.09	-1.78	-1.43	2904	1
Intercept[3]	-0.41	0.09	-0.58	-0.24	2838	1
Intercept[4]	0.25	0.09	0.08	0.42	2836	1
Intercept[5]	1.32	0.09	1.15	1.49	2927	1
Intercept[6]	2.7	0.09	2.52	2.88	3101	1
MusSoph	-0.13	0.08	-0.29	0.03	2586	1
TrialNo	0.03	0.04	-0.05	0.10	10652	1
lnContTrialNo	-0.03	0.02	-0.08	0.02	17622	1
TrialNo <sup>2</sup>	-0.04	0.04	-0.12	0.05	8333	1
ChromSim	-0.01	0.05	-0.11	0.09	8150	1
DiatSim	-0.02	0.06	-0.14	0.09	7195	1
Height	0.23	0.19	-0.14	0.60	5280	1
RelHeight	-0.12	0.18	-0.49	0.23	5473	1
Height <sup>2</sup>	-0.04	0.10	-0.23	0.14	5631	1
RelHeight <sup>2</sup>	-0.08	0.10	-0.27	0.11	5629	1

Continued on next page

Table D.5 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
Primacy	-0.05	0.10	-0.24	0.14	12973	1
Previous	0.24	0.05	0.14	0.34	5745	1
Recency	0.32	0.13	0.08	0.57	10074	1
SPCS	0.44	0.06	0.32	0.56	5276	1
MelCont	-0.02	0.03	-0.08	0.03	13517	1
Count	0.02	0.04	-0.06	0.10	11314	1
ModeHeight	0.03	0.03	-0.03	0.10	11508	1
DiatBoost	0.04	0.04	-0.03	0.11	10233	1
MusSoph:Height	0.19	0.11	-0.03	0.42	7206	1
MusSoph:RelHeight	-0.19	0.11	-0.41	0.02	8422	1
MusSoph:Height <sup>2</sup>	-0.01	0.06	-0.12	0.10	7995	1
MusSoph:RelHeight <sup>2</sup>	0.06	0.06	-0.05	0.17	8261	1
MusSoph:Primacy	0.00	0.07	-0.14	0.14	16240	1
MusSoph:Previous	0.01	0.05	-0.09	0.10	5711	1
MusSoph:Recency	0.38	0.11	0.17	0.60	9721	1
MusSoph:SPCS	0.26	0.05	0.16	0.37	5682	1
MusSoph:MelCont	-0.01	0.02	-0.04	0.03	21218	1
MusSoph:Count	0.04	0.03	-0.02	0.10	9752	1
MusSoph:ModeHeight	0.08	0.02	0.04	0.13	13014	1
MusSoph:DiatBoost	0.08	0.03	0.03	0.14	9209	1
TrialNo:Height	-0.11	0.10	-0.31	0.09	5869	1
TrialNo:RelHeight	0.06	0.10	-0.14	0.26	6401	1
TrialNo:Height <sup>2</sup>	0.00	0.05	-0.1	0.11	10082	1
TrialNo:RelHeight <sup>2</sup>	-0.02	0.05	-0.12	0.09	10236	1
TrialNo:Primacy	0.11	0.07	-0.03	0.24	16389	1
TrialNo:Previous	-0.03	0.02	-0.07	0.00	17672	1
TrialNo:Recency	-0.1	0.07	-0.25	0.04	13718	1
TrialNo:SPCS	-0.04	0.03	-0.09	0.01	10578	1
TrialNo:MelCont	0.00	0.02	-0.04	0.04	18620	1
TrialNo:Count	0.07	0.05	-0.02	0.16	9109	1
TrialNo:ModeHeight	0.02	0.02	-0.03	0.07	10681	1
TrialNo:DiatBoost	0.05	0.02	0.01	0.09	14452	1
InContTrialNo:Height	-0.03	0.08	-0.18	0.13	10950	1
InContTrialNo:RelHeight	0.02	0.08	-0.13	0.18	10984	1
InContTrialNo:Height <sup>2</sup>	0.03	0.04	-0.06	0.12	10568	1
InContTrialNo:RelHeight <sup>2</sup>	-0.04	0.05	-0.13	0.05	10551	1
InContTrialNo:Primacy	0.02	0.07	-0.11	0.15	18676	1
InContTrialNo:Previous	0.02	0.02	-0.02	0.06	15234	1
InContTrialNo:Recency	0.03	0.06	-0.1	0.16	16638	1
InContTrialNo:SPCS	0.04	0.02	0.00	0.08	14209	1
InContTrialNo:MelCont	0.02	0.02	-0.02	0.06	19878	1
InContTrialNo:Count	0.03	0.02	-0.01	0.07	16461	1
InContTrialNo:ModeHeight	0.02	0.02	-0.02	0.06	13367	1
InContTrialNo:DiatBoost	0.01	0.02	-0.02	0.05	19605	1
TrialNo <sup>2</sup> :Height	-0.2	0.14	-0.47	0.07	4826	1
TrialNo <sup>2</sup> :RelHeight	0.21	0.14	-0.06	0.48	4996	1
TrialNo <sup>2</sup> :IHeight <sup>2</sup>	-0.04	0.07	-0.18	0.10	5702	1
TrialNo <sup>2</sup> :IRelHeight <sup>2</sup>	0.03	0.07	-0.11	0.17	5652	1
TrialNo <sup>2</sup> :Primacy	-0.03	0.07	-0.17	0.12	10694	1
TrialNo <sup>2</sup> :Previous	0.00	0.03	-0.05	0.05	11398	1
TrialNo <sup>2</sup> :Recency	0.13	0.09	-0.05	0.30	12119	1
TrialNo <sup>2</sup> :SPCS	0.02	0.03	-0.03	0.07	12385	1

Continued on next page

Table D.5 – Continued from previous page

Effect	Estimate	Est.Error	l-95% CI	u-95% CI	Eff.Sample	Rhat
TrialNo <sup>2</sup> :MelCont	0.02	0.02	-0.02	0.06	12823	1
TrialNo <sup>2</sup> :Count	-0.05	0.03	-0.11	0.01	9438	1
TrialNo <sup>2</sup> :ModeHeight	0.04	0.02	0.00	0.09	10738	1
TrialNo <sup>2</sup> :DiatBoost	-0.02	0.02	-0.06	0.03	11993	1
ChromSim:Height	0.13	0.23	-0.33	0.59	5277	1
ChromSim:RelHeight	-0.15	0.24	-0.62	0.31	5276	1
ChromSim:Height <sup>2</sup>	0.07	0.12	-0.16	0.29	6568	1
ChromSim:RelHeight <sup>2</sup>	-0.06	0.12	-0.29	0.18	6531	1
ChromSim:Primacy	-0.15	0.13	-0.39	0.10	11494	1
ChromSim:Previous	0.11	0.04	0.04	0.18	11548	1
ChromSim:Recency	-0.02	0.13	-0.28	0.23	10747	1
ChromSim:SPCS	-0.05	0.04	-0.13	0.03	10116	1
ChromSim:MelCont	-0.02	0.03	-0.09	0.04	10624	1
ChromSim:Count	-0.02	0.04	-0.11	0.06	9091	1
ChromSim:ModeHeight	0.05	0.05	-0.05	0.13	12426	1
ChromSim:DiatBoost	0.00	0.05	-0.09	0.10	11828	1
DiatSim:Height	-0.14	0.22	-0.56	0.28	4756	1
DiatSim:RelHeight	0.17	0.22	-0.25	0.60	4738	1
DiatSim:Height <sup>2</sup>	-0.09	0.11	-0.3	0.12	6026	1
DiatSim:RelHeight <sup>2</sup>	0.06	0.11	-0.15	0.27	5981	1
DiatSim:Primacy	0.12	0.13	-0.13	0.37	11008	1
DiatSim:Previous	-0.1	0.04	-0.18	-0.03	10868	1
DiatSim:Recency	-0.17	0.14	-0.45	0.10	10590	1
DiatSim:SPCS	0.11	0.05	0.02	0.20	9361	1
DiatSim:MelCont	0.00	0.03	-0.07	0.07	11210	1
DiatSim:Count	-0.03	0.04	-0.11	0.04	9591	1
DiatSim:ModeHeight	0.03	0.04	-0.04	0.10	12121	1
DiatSim:DiatBoost	0.04	0.05	-0.06	0.13	11300	1