Patterns of individual differences in the perception of missing-fundamental tones

D. Robert Ladd[1], Rory Turnbull[1,4], Charlotte Browne[1], Catherine Caldwell-Harris[3], Lesya Ganushchak[2], Kate Swoboda[1,5], Verity Woodfield[1], Dan Dediu[2]

[1]School of Philosophy, Psychology and Language Sciences, University of Edinburgh
[2]Max-Planck Institute for Psycholinguistics
[3]Department of Psychology, Boston University
[4]Department of Linguistics, Ohio State University
[5]School of Life and Health Sciences at Aston University, Birmingham

In press, *Journal of Experimental Psychology: Human Perception and Performance*

Author Note: Thanks to Eddie Dubourg and Simon Kirby for their technical contributions and to Richard Shillcock, Morten Christiansen, Tim Bates and Antje Meyer for discussion. The pilot experiments, which determined the general direction of subsequent work, were carried out jointly by DRL, RT and DD. The other authors were each involved in one of the four larger studies: CC-H in Exp. 3b, LG in Exp. 4a, KS in Exp. 4b, and CB and VW in Exp. 4c. Expt. 4c was part of two student papers. DRL prepared the stimuli for all experiments and had primary responsibility for writing the paper, while DD had primary responsibility for the statistical analyses.

Correspondence concerning this article should be addressed to Prof. D. Robert Ladd, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh EH8 9AD, Scotland. Email: bob.ladd@ed.ac.uk

Abstract: Recent experimental findings suggest stable individual differences in the perception of auditory stimuli with missing fundamental frequency (F0). Specifically, some individuals readily identify the pitch of such tones with the missing F0 ('F0 listeners'), and some base their judgement on the frequency of the partials that make up the tones ('spectral listeners'). However, the diversity of goals and methods in recent research makes it difficult to draw clear conclusions about individual differences. The first purpose of this paper is to discuss the influence of methodological choices on listeners' responses. The second goal is to report findings on individual differences in our own studies of the missing-fundamental phenomenon. We conclude that there are genuine, stable individual differences underlying the diverse findings, but also that there are more than two general types of listeners, and that stimulus variables strongly affect some listeners' responses. This suggests that it is generally misleading to classify individuals as 'F0 listeners' or 'spectral listeners'. It may be more accurate to speak of two *modes of perception* ('F0 listening' and 'spectral listening'), both of which are available to many listeners. The individual differences lie in what conditions the choice between the two modes.

Keywords: missing fundamental, pitch perception, individual differences

November 2012

A missing fundamental (MF) tone is an artificially constructed acoustic stimulus consisting of a number of component frequencies, chosen so that they could be the harmonics of some fundamental frequency (F0) that is itself not present in the stimulus. For example, consider a tone consisting of energy at 750 Hz, 1000 Hz, and 1250 Hz. The lowest common factor of these frequencies is 250 Hz, and in general such a tone is often perceived as having a pitch of 250 Hz; that is, the pitch percept may be based on a frequency that is in some sense not physically present in the stimulus. This frequency – also referred to in the literature as 'virtual pitch' (e.g. Terhardt, 1979), 'periodicity pitch' (e.g. Licklider, 1951), and 'residue pitch' (e.g. Schouten, 1940) – is the missing fundamental. However, it is also possible to perceive the MF tone just described as a chord consisting of the component frequencies (the 'partials') that are actually present in the stimulus: specifically, in musical terms, as an inverted major triad (roughly a very flat G_5 C_6 E_6 [g″ c‴ e‴]). The starting point for this paper is the finding that many individuals seem to have stable biases in the way they perceive MF stimuli, preferentially hearing the pitch of the stimulus either on the basis of the MF or of the partials that are actually present.

The source of these individual differences is not known. Recent interest in this topic has arisen within cognitive neuroscience, especially among those interested in music perception and cognition. Some of this work seeks to correlate different patterns of responses to MF stimuli with neuroanatomical (e.g. Schneider et al., 2005) or neurophysiological (e.g. Patel & Balaban, 2001) differences; other work emphasises the influence of experience, particularly musical training, on the patterns of perceptual responses (e.g. Seither-Preisler et al., 2007). However, it is also known that there are purely physical effects that influence the actual acoustic nature of signals consisting of a small number of partials, and that these may affect the cochlear response to the signals; probably the most important effect of this sort is the existence of 'combination tones' (see e.g. Terhardt, 1974; Moore, 2012). There is a separate line of recent research on MF perception among hearing researchers that seeks to understand these basic physical mechanisms (e.g. Bernstein & Oxenham, 2006; Gockel, Plack, & Carlyon, 2005; Gockel, Carlyon, & Plack, 2010; see Moore & Gockel, 2011 for a recent review). It is entirely possible that some of the individual differences under discussion here are based on different cochlear responses to differences in the signal, rather than originating in the brain.

However, the present paper is concerned not with the basis of the behavioural differences, but with a clearer definition of the differences themselves. Recent work is extremely diverse methodologically, and has focused on testing hypotheses about the effect of specific individual differences (e.g. differences of musical training) on the perception of MF stimuli. Moreover, it has tended to proceed as if the behavioural differences are straightforwardly binary, describing individuals as belonging to one of two basic types of listeners. Our investigations have shown that this approach oversimplifies the nature of the individual differences, and we believe that this oversimplification directly affects our ability to look for their underlying causes. Our aim in this paper is to present a more refined characterisation of the behavioural differences, which will be of use to subsequent research on any aspect of the MF phenomenon.

The missing-fundamental task

Basic design

The first systematic exploration of individual differences in responses to MF tones was carried out by Smoorenburg (1970), who seems to have stumbled on the existence of the individual differences while researching the basic physics of the phenomenon (1970, p. 927). Smoorenburg developed an ostensibly simple way to determine whether a listener is taking the missing F0 or one of the partials as the pitch of a MF tone. By presenting MF tones in pairs, he was able to construct stimuli that would appear to go either up or down in pitch from the first member of the pair to the second, depending on whether the pitch of the individual members of the pair was being perceived on the basis of the MF or of the partials. This behavioural task is the experimental tool on which subsequent research has been based.

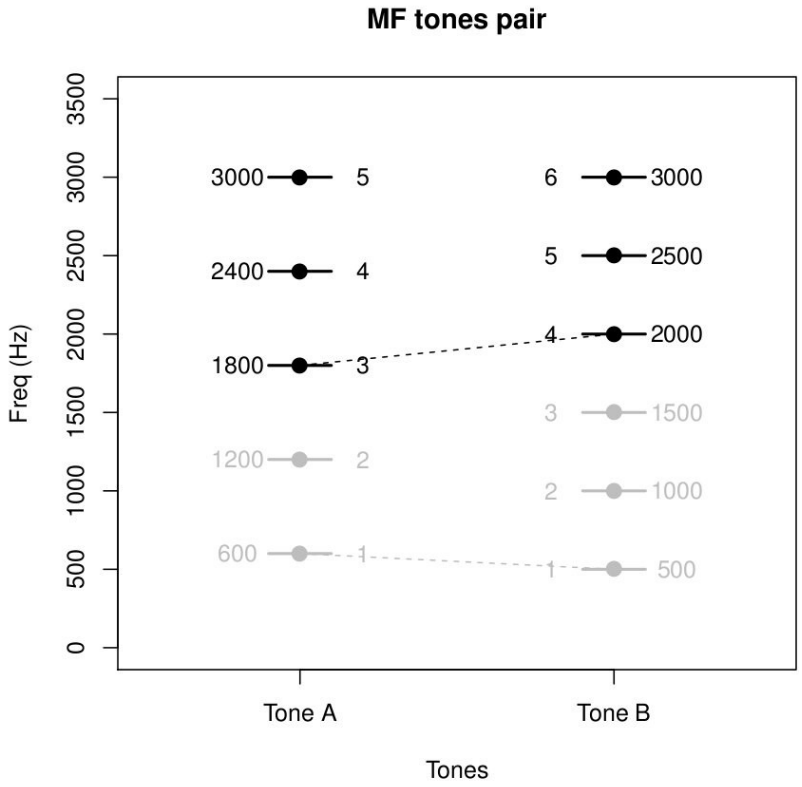

Figure 1. Basic design of MF task stimuli. Tone A (on the left) consists of three partials that could be the third, fourth and fifth harmonic of a fundamental frequency (the 'first harmonic') that is not physically present in the signal. Tone B (on the right) also consists of three partials, which could be the fourth, fifth and sixth harmonics of a fundamental frequency (also not physically present). Crucially, the missing fundamental in Tone B is lower than that in Tone A, while the lowest frequency actually present in Tone B is higher than the lowest frequency actually present in Tone A.

The basic design of stimuli in the MF task is diagrammed in Fig. 1. In this example, it can be seen that the MF 'goes down' (i.e. is lower in Tone B than in Tone A), but the

lowest partial actually present in the stimuli 'goes up' (i.e. is lower in Tone A than in Tone B). This ambiguity can be achieved even while keeping the highest partials at the same frequency in both tones; all that is needed is to treat that top frequency as the n^{th} harmonic in Tone A and the $(n+1)^{th}$ harmonic in Tone B. To avoid misunderstanding, it is worth mentioning that the terms 'Tone A' and 'Tone B' are used only for clarity of reference and imply nothing about order of presentation. In the various studies discussed here, actual stimulus pairs were of course presented in either order (AB or BA), or in both orders. No source reports any order effects, but as we shall see, such effects do occur, which complicates the interpretation of what listeners are actually doing in the MF task.

Individual differences in MF perception

Smoorenburg's experiment suggested that most individuals fairly consistently perceive the pitch of the MF tones either in terms of the missing F0 or on the basis of the component frequencies. His data also made it appear that there are roughly equal numbers of the two types of listeners. However, his procedure involved only two different stimuli, presented repeatedly (i.e. a single pair of 'Tone A' and 'Tone B' in both orders of presentation). If there is a genuine source of individual difference, it is not surprising that this procedure would lead to a strong separation of the two response patterns. More recent studies seem to show that if listeners are presented with a range of different MF stimuli, their behaviour may be more variable, and that the properties of the stimulus may have consistent influences on which way listeners tend to hear it.[1] We focus here on a comparison of two large studies, Schneider et al. (2005) and Seither-Preisler et al. (2007).

Schneider et al. (2005) was a large study of musicians and non-musicians, with the primary aim of relating differences in MF perception to differences in neuroanatomy, specifically to differences in the volume of the pitch-detection areas in left and right Heschl's gyrus. A secondary aim was to explore the effect of certain stimulus variables (e.g. number of partials present in the stimulus tones) on the perception of MF tones. Schneider et al. reduced listeners' overall pattern of responses to a quotient whose value ranges from –1 to +1, according to the proportion of responses based on F0 and on the partials. They report a bimodal (broadly U-shaped) distribution in the value of this quotient, with a minimum in the middle of the range (around 0, where an individual's responses are mixed). On this basis they divide the range in half and classify listeners as 'F0 listeners' or 'spectral listeners'. We will adopt this terminology here.[2] Schneider et al. also report that, on average, spectral listeners have greater cortical volume in Heschl's gyrus in the right hemisphere than in the left, while F0 listeners have greater volume in the left than in the right. They found no consistent difference in responses or in hemispheric asymmetry between musicians and non-musicians, but report overall larger Heschl's gyrus volume in musicians.

---

[1] Louis Pols (personal communication, September 2011) tells us that he worked in the same lab as Smoorenburg at the time of the experiments on which the 1970 paper was based, and says that Smoorenburg was well aware that some MF tones would elicit F0 percepts from most listeners. Stimuli had to be carefully chosen in order to draw out the difference between individuals.

[2] 'Synthetic' and 'analytic' are two common terms used for F0 and spectral listeners respectively, and are widely used in the literature (e.g. Schneider & Wengenroth, 2009). Although this pair of terms has a long history (Houtsma & Fleuren, 1991 attribute the terms to Helmholtz), we prefer the terms from Schneider et al. 2005, which are more theoretically neutral.

Related work by Schneider and Wengenroth (2009) suggests that there may be differences among musicians depending on their instrument or the type of music they play, e.g. that jazz musicians are more likely to be spectral listeners than classical musicians.

Seither-Preisler et al. (2007) also studied musicians and non-musicians; they did not do any brain imaging, but their hypotheses are implicitly driven by assumptions about brain plasticity, specifically the effect of musical training. Like Schneider et al., their materials manipulated a number of different stimulus variables, but the variables they explored differed quite considerably from those studied by Schneider et al. They also used very different (and more complex) statistical reductions of individuals' behavioural response patterns that ultimately abstracted away from the effect of stimulus variables. Like Schneider et al., they found that many participants responded as F0 listeners or spectral listeners, and indeed, they report a sharper dichotomy between the two groups than was found by Schneider et al. However, this sharper dichotomy is due in part to their analysis procedures, which led them to exclude roughly a quarter of their participants on the grounds that their responses were not reliably distinguishable from guesswork. They also, unlike Schneider et al., showed a clear effect of musical training, with professional musicians responding far more often as F0 listeners. Note in this connection that Seither-Preisler et al.'s repeated references to 'guessing' may seem to suggest that there is a right answer (viz., F0 response), an implication that we find unjustified.

Stimulus variables in the MF task

One of the striking features of the two studies just summarised is that they make very different methodological choices in their procedures and in constructing their stimuli, yet both find evidence for Smoorenburg's basic conclusion that listeners exhibit two essentially different types of behaviour in processing MF stimuli. Other recent studies, based on still other methodological approaches, lead to the same conclusion. For example, Patel and Balaban (2001), a study of neural activity in pitch perception with a focus on the relation between time-domain and frequency-domain processing, also finds clear evidence that individuals tend to favour one of two different modes of behaviour. The fact that these differences show up in a wide variety of experimental situations suggests that the underlying phenomenon is very robust.

At the same time, early psychoacoustic work into the nature of MF perception in general (Plomp, 1967; Ritsma, 1962, 1963a, 1963b) has demonstrated that stimulus properties can have consistent effects on listeners' responses. Systematic manipulation of stimulus variables in subsequent work (e.g. Moore, Glasberg, & Peters, 1985; Houtsma & Fleuren 1991) further established the role of stimulus properties in determining response patterns, independent of individual differences. These effects were not absent from Schneider et al. and Seither-Preisler et al.'s results. Two such findings emerge clearly from these two papers:

- As the musical interval between the missing fundamentals in Tone A and Tone B increases, listeners are more likely to give F0 responses. This effect was demonstrated clearly by Seither-Preisler et al. (their Fig. 3, p. 746, cf. Meddis & Hewitt, 1991; Moore et al., 1985).
- As the number of partials in the tones increases, listeners are also more likely to base their pitch judgement on the missing F0. This effect was

systematically shown by Schneider et al. (their Fig. 1d, p. 1242; cf. Faulkner, 1985; Ritsma, 1962).

This means that, irrespective of an individual's bias toward F0 or spectral listening, responses can be influenced by differences of detail in the stimuli. It therefore seems important to consider methodological choices in stimulus construction more closely. Unfortunately, this is not as straightforward as it might sound, because the stimulus variables are highly interdependent. We cannot simply vary them orthogonally to explore their effects.

This interdependency can be illustrated clearly by the relation among what we might refer to as top frequency (the frequency of the highest partial), harmonic rank (the position of the partials in the harmonic series, e.g. 5^{th} and 6^{th} harmonics), and the interval between the missing F0 of Tone A and Tone B. If top frequency is held constant within a stimulus pair (as was done by Schneider et al.), then interval is completely determined by the choice of harmonic rank for the two stimulus tones (or vice-versa); if interval is systematically varied (as was done by Seither-Preisler et al.), then the top frequency of the two stimulus tones is completely determined by their harmonic rank (or vice-versa). For example, if the top frequency of a stimulus pair is kept constant at 600 Hz and we specify the top partials in tones A and B as having harmonic rank 5 and 6 respectively – which corresponds roughly to the procedure of Schneider et al. – the interval will necessarily be a minor third (3 semitones), because the ratio of the virtual F0 of the two stimulus tones will be 6:5 (120 Hz and 100 Hz). If the top frequency is held constant at 600 Hz and we want to specify an interval of a fifth (ratio 3:2), we would have to use harmonic rank 4 and 6 (or 6 and 9, or 8 and 12, etc.). Conversely, if we specify an interval of a fifth and also specify the harmonic rank of tone A and B – which corresponds roughly to the procedure of Seither-Preisler et al. – then the top frequency of one stimulus tone will necessarily be higher than the other. Similar interdependencies affect other stimulus variables; fuller discussion is beyond the scope of this report.

This interdependency makes it difficult to interpret some of the findings reported in the papers under consideration, or to investigate apparent contradictions. The most obvious discrepancy here involves overall frequency level and harmonic rank. Schneider et al. report an effect of 'average spectral frequency' (their Fig. 1c, p. 1242): as the average frequency of the stimulus tones increases, so too (albeit rather irregularly) does the number of F0 responses. At the same time, they also report an effect of harmonic rank, such that partials *lower* in the harmonic series evoke more F0 responses (their Fig. 1d, p. 1242); Seither-Preisler et al. (p. 745f) mention a similar effect of harmonic rank in a variable they call 'spectral profile'. These findings make exactly opposite predictions about the effect of manipulating the partials in a MF tone pair at a given F0 level: higher partials will raise the average spectral frequency and therefore should lead to more F0 responses, yet higher partials will also be higher in the harmonic series and therefore should lead to more spectral responses. Furthermore, in a pair of MF tones constructed according to Schneider et al.'s procedures, higher partials will yield smaller intervals between the missing F0 of the two tones, which (given Seither-Preisler et al.'s results) should lead to more spectral responses as well. Since it is physically impossible to vary harmonic rank, MF interval, and average spectral frequency orthogonally while keeping F0 within a

constrained range, we cannot resolve these contradictory predictions in conventional experimental ways.

Classification of listeners

Given the forced-choice approach of the experiments just discussed, labels such as 'F0 listening' and 'spectral listening' can certainly be applied to individual *responses*. However, it is less clear that these labels can also be appropriately used to describe the overall behaviour of *listeners* – that is, whether individuals clearly fall into two groups with distinct behavioural strategies. It seems likely that there really are distinct behavioural strategies, but the matter is not simple, and it depends to some extent on how we quantify overall patterns of individual responses.

Schneider et al. add each participant's responses together and compute an individual 'index' that expresses the proportion of F0 and spectral responses on a scale from –1 to +1. We refer to this score in what follows as the Schneider Index (SI). Their formula is as follows:

$$\mathrm{SI} = \frac{\mathrm{sp} - \mathrm{f0}}{\mathrm{sp} + \mathrm{f0}} \tag{1}$$

where f0 refers to the number of F0 responses and sp refers to the number of spectral responses. Seither-Preisler et al. use a similar score to describe individual performance on their Auditory Ambiguity Test (AAT), which simply reports the overall proportion of F0 responses on a scale from 0 to 1.0. These two measures are completely equivalent, with SI of –1 corresponding to 1.0 on the AAT, SI of +1 corresponding to 0, and SI of 0 corresponding to 0.5.[3] As noted above, both teams report bimodal distributions of these quantitative measures, with many listeners having scores near the ends of the range and fewer in the middle.

The most important problem with this approach to data reduction is intra-individual consistency. Some participants give completely consistent responses – that is, 100% of their responses are either 'F0' or 'spectral'. In these cases, there is no issue about describing individuals as 'F0 listeners' or 'spectral listeners'. However, many participants give a mix of responses, which can yield SI near 0. It is not immediately obvious how to treat such mixed behaviour.

Schneider et al. hypothesised that some degree of inconsistency might arise through what they called octave-shifting, i.e. perceiving the *second* harmonic (one octave higher than the missing F0) as the pitch of a MF tone. They attempted to allow for this kind of inconsistency by including control stimuli in which Tone A actually includes the F0 (in terms of the example shown in Figure 1, Tone A would have

[3] There is unfortunately a discrepancy between the formula given on p. 1242 of Schneider et al.'s paper and the published graphs in the same paper: in the formula, F0 responses are positively poled (i.e.100% F0 responses yields a SI of +1) while in the graphs, F0 responses are negatively poled (i.e. 100% F0 responses yields a SI of –1). Subsequent work by Schneider and his colleagues (e.g. Schneider and Wengenroth, 2009) has settled on the polarity shown in the graphs, and this is reflected in the formula we use here. Note, though, that this is in some sense opposite to the polarity implicit in Seither-Preisler et al.'s AAT. Ultimately, of course, the choice is arbitrary, and for exactly that reason there is considerable potential for confusion. *Caveat lector*.

included partials at 1200 and 600 Hz in addition to the higher harmonics). In such a stimulus, a listener who was truly perceiving the MF as the pitch of Tone B would respond 'down', but a listener who was perceiving the second harmonic would respond 'up'. Schneider et al. excluded such octave-shifted responses from their analysis altogether, calculating SI only on the basis of responses that could be clearly classed as F0 or spectral. In keeping with the importance of stimulus variables discussed in the preceding section, Schneider et al. note that octave-shifted responses were given primarily to stimuli with relative high MF values.

Seither-Preisler et al. took a different approach to inconsistent responses; as noted above, they simply excluded many participants whose AAT scores fall in the middle of the range on the grounds that such response patterns cannot be distinguished from guesswork. At the same time, they suggest that such mid-range scores might arise for two distinct reasons: either the participants are responding *inconsistently* (that is, giving opposite responses to different presentations of the same stimulus), or they are responding *inhomogeneously* (that is, consistently giving F0 responses to some stimuli and spectral responses to others). This is a valuable distinction, especially in light of the clear findings, summarised above, that certain stimulus variables systematically influence the overall proportion of F0 responses, and in light of Schneider et al.'s finding that some stimulus types seem to yield more octave-shifted percepts. If many individuals exhibit systematic inhomogeneous behaviour, then it is obviously an oversimplification to describe everyone as either a spectral listener or an F0 listener.

However, Seither-Preisler et al. were limited in their ability to detect inhomogeneity directly, because their participants heard only a few presentations of each of many stimulus types. Consequently, some of the participants with mid-range AAT scores who were excluded for inconsistency might more appropriately have been treated as inhomogeneous. Furthermore, the very notions of inconsistency and inhomogeneity are based on an easily overlooked assumption underlying the MF task itself. Despite its apparent simplicity, the task presupposes that listeners' responses reflect *independent* percepts of the pitch of the two tones in each stimulus. That is, it assumes that listeners perceive the pitch of Tone A and Tone B according to either the MF or the partials, and report a pitch rise or fall across the stimulus on that basis. It does not allow for the possibility that listeners who are asked to report the *direction* of pitch across the stimulus do so on some more holistic basis that does not simply reflect how they perceive static pitch in a single tone (cf. the discussion of contour and interval in Patel 2008, chapter 4); as we shall see, there is reason to think that this possibility must be taken seriously. In any case, one of the central goals of the work reported here is a better understanding of response patterns that yield intermediate values of SI.

Our studies

Our studies of this topic are ultimately motivated by an interest in individual perceptual and cognitive differences that are potentially relevant to language. However, the focus of the present paper is more basic. In order to draw convincing connections between specific individual behavioural differences and other cognitive traits, we will need a well-understood and well-operationalised measure of the behaviour in question. As can be seen from the foregoing review, this is precisely what we do not have in the case of the MF task. What we report here is therefore a

set of experiments aimed primarily at clarifying what it is that the MF task reveals. Our principal concern is with the distinction between inconsistency and inhomogeneity, and with explanations for mid-range SI scores. We also report findings on test-retest reliability, and in a limited way we deal with the related issue of the effects of stimulus variables, discussed above. In keeping with our ultimate interest in individual differences, we also report findings on the influence of three participant variables, namely age, gender, and musical background.

The data reported in Experiments 1, 2, and 3a come from strictly exploratory experiments. The remaining data, in Experiments 3b, 4a, 4b, and 4c, are drawn from four studies that focused on the relation between the MF task and other perceptual measures relevant to language. The specific issues addressed in the last four experiments have been (or will be) reported elsewhere, and the present paper includes only the basic behavioural data from those experiments. Although the studies had different purposes, the same task is used and methodologies are broadly similar. Most importantly, the minor methodological differences in our experiments had no impact on the conclusion that there are two different ways of responding to MF stimuli; indeed, as we have already discussed, experiments in the literature have diverged radically in their methodological choices yet have all converged on this conclusion. It thus made sense to pool the data across the studies given the benefits of increasing generalizability and statistical power. Detailed discussion of the comparability of the different experiments is provided in the online appendix.

Method

Stimulus variables

By and large our approach to stimulus construction was closer to that of Schneider et al. than to that of Seither-Preisler et al. Within a stimulus we always held the top frequency of Tone A and Tone B constant, and always kept the harmonic rank of the two sets of partials close (that is, our stimuli resemble the one illustrated in Fig. 1). This in turn means that the interval between the two missing F0 values was always quite small, between 2 and 4 semitones. It also means that the F0 value of the tones was determined entirely by the top frequency and the harmonic rank of the partials, and that the range of F0 values was therefore, especially in our earlier experiments, quite large. In the later experiments (Experiments 4a, 4b and 4c), influenced by Seither-Preisler et al., we narrowed the range of top frequencies and used lower harmonic ranks, thereby narrowing the range of the missing F0. In the earlier experiments the tones consisted of three partials, but in the Experiment 4 set we used a mix of two-partial and three-partial stimuli.

The most significant respect in which our work diverges methodologically from that of Seither-Preisler et al. and especially Schneider et al. is that our experiments involve *fewer stimulus types and more responses to each type*. For example, in Experiment 1, we had only 15 stimulus types, based on a two-dimensional stimulus matrix with 5 settings of the top frequency and 3 settings of the harmonic rank of the partials. In each of the 15 cells of this stimulus matrix, every participant gave 10 judgements during the course of the experiment, five in each order of presentation (AB or BA). By comparison, Seither-Preisler et al. had 50 stimulus types, and participants gave only 4 judgements per stimulus type, two in each order. Schneider et al. had 144

stimulus types and obtained only one response per type; the order of Tone A and Tone B within each stimulus type was randomly assigned. By contrast, all of our data (with minor exceptions due to errors and missing responses, and with the systematic exception of Experiment 4b) are based on 10 responses per stimulus type. This gives us a good basis for investigating Seither-Preisler et al.'s distinction between inhomogeneity and inconsistency in participants whose responses are not consistently at one end or the other of the SI scale. Quite unexpectedly, it also allowed us to observe a large order-of-presentation effect in some participants, reported in more detail below, which we believe is relevant to the interpretation of intermediate values of SI.

Table 1
Summary of experiments.

Expt.	No. of participants	Stimulus matrix	
		top frequencies	spectral composition
1	37 (16 F, 21 M)	300, 500, 900, 1400, 2200 Hz	345/456, 567/678, 689/890
2	20 (12 F, 8 M)	500, 750, 1050, 1400, 1800 Hz	345/456, 678/789
3a[a]	23 (13 F, 10 M)	250, 750, 1050, 1400, 1800, 2000, 2500, 3000, 4000, 5000, 6000 Hz	345/456, 678/789
		300, 500, 900, 1400, 2200 Hz	567/678
3b[b]	152 (105 F, 47 M)	[as 3a]	[as 3a]
4a[c]	50 (39 F, 11 M)	500, 675, 900, 1200, 1600, 2150 Hz	**34/45**, 345/456, **56/67**
4b[d]	73 (57 F, 16 M)	[as 4a]	[as 4a]
4c[e]	57 (41 F, 16 M)	[as 4a]	[as 4a]

Note. Under 'stimulus matrix', frequency level is indicated by the top frequency of the stimulus, while spectral composition is indicated by the harmonic rank of all partials in both stimulus tones. The harmonic rank of the partials is specified in abbreviated form as e.g. 345/456, which is to be read as meaning that Tone A consists of harmonics 3, 4 and 5 of the MF, while Tone B consists of harmonics 4, 5 and 6. In these abbreviated formulas, harmonic 10 (which was used only in Experiment 1) is symbolised as 0. In the Experiment 4 set, two of the spectral composition conditions (in boldface) involve only two partials. [a] *Participants were amateur choral singers. Intended to explore findings about musical preferences in Schneider and Wengenroth (2009), but inconclusive in that respect.* [b] *Part of a study on individual differences in a task involving implicit learning of an artificial tone language (Caldwell-Harris, Biller, Ladd, Dediu, & Christiansen, 2012).* [c] *Part of a large study on individual differences, with both language-related (e.g. non-word repetition, vocabulary learning) and control tasks (e.g. IQ). Includes test-retest reliability data.* [d] *Part of a study on hemispheric differences in pitch processing, to be reported in a separate paper.* [e] *Part of two separate studies, one on native language and MF perception, and one exploring a possible link between MF perception and the 'tritone paradox' (Deutsch 1991, Repp 1994).*

Summary of experiments

Table 1 summarises the details of all the experiments on which this report is based. All are based on a systematic two-dimensional matrix of stimulus types like the one just exemplified for Experiment 1, with *top frequency* and *spectral composition* (here

defined as the harmonic rank of the partials) as the two dimensions of the matrix. The number of stimulus types (i.e. the number of cells in the matrix) varies from 10 to 27. Experiments whose identifiers share a number (e.g. 3a and 3b) have the same stimulus matrix but are not otherwise related. As noted above, most of the experiments were motivated by research questions beyond the basic goal of clarifying the nature of response patterns in the MF task. These additional questions are summarised in the notes to Table 1.

Participants

Altogether there were 412 participants. In Experiments 1 and 2 they were mostly friends, colleagues, or family members of one or more of the authors; in Experiment 3a they were amateur singers from two respected Edinburgh choirs; in Experiment 3b they were mostly students at Boston University; in Experiment 4a they were mostly students at the University of Nijmegen; in Experiments 4b and 4c they were mostly students at the University of Edinburgh. In the experiments involving students, the participants were paid a small sum for participating; in Experiment 3a a small donation was made to the choirs in which the participants sang. Overall, the great majority of participants were native speakers of English, but there were also native speakers of quite a few other languages as well, in particular Dutch and Chinese. Except in Experiments 3b and 4c, native language or language background was not a variable we were interested in, or one we attempted to control; except in those two experiments, participants were almost all white Europeans or European-Americans, and even in those two experiments the majority of participants were white native speakers of English or another European language. Participants' ages varied from about 15 to about 75, but in all the experiments that relied on students as participants (3b, 4a, 4b and 4c) most were in their early 20s. In Experiments 1 and 2 ages were more widely distributed, while in the experiment involving choral singers (Experiment 3a) most of the participants were middle-aged or older (median age 55).

Stimulus preparation

We created our stimuli using an application written for us by Simon Kirby, based on Max/MSP software, which allowed us to specify (a) the F0 and duration of the two tones, (b) the number, harmonic rank, and relative amplitude of the partials, and (c) the duration of the gap between the tones. Except in the second phase of Experiment 4c (see 'Procedure', below), the two tones in each stimulus were 500 ms long, with a 250 ms gap between them, corresponding exactly to Schneider et al.'s stimuli (Seither-Preisler et al. used tones 500 ms long with a 500 ms gap). All our stimulus tones had flat spectra, again following Schneider et al. rather than Seither-Preisler et al.; we experimented informally with modifying the spectral slope and concluded that except in cases of very steep slope there was no readily perceptible difference, but we have not done a controlled comparison. Most of the experiments were done with a version of the software that did not control the phase relations of the component partials; Experiment 4c used a newer version in which phase can be controlled, and the stimuli were created with the partials in phase. Note in this connection that Smoorenburg specifies that the tones in his stimuli were not phase-controlled; Seither-Preisler et al. do not specify; Schneider (personal communication, August 2012) reports that the stimuli in Schneider et al. 2005 were phase-controlled. Sound files of an illustrative sample of the stimuli can be found in the online appendix.

Procedure

Most of the experiments were run using an e-prime script written for us by Eddie Dubourg, but Experiments 2 and 4a used a Presentation® software (www.neurobs.com) script written by DD. There were minor variations in the instructions given to participants, but the instructions were always presented in writing on screen at the beginning of the experiment, and they always involved an explanation that the stimuli consisted of two tones in sequence and that the experimental task was to judge whether the pitch went up or down from the first to the second. The instructions also made clear that it might be difficult to judge and told participants to give their dominant impression. When the stimulus was played the screen displayed a prompt for a response; participants responded by pressing one of two keys on the keyboard, e.g. U for 'up' and D for 'down'. In Experiment 1 the response trials timed out after 2 seconds; subsequently the program was set up so that the participant had to give a response before the next stimulus was presented. All experiments included a short practice session.

Listening conditions varied somewhat: in Experiment 1 some of the participants used a laptop with ordinary headphones or (in a few cases) the laptop's internal speakers, but all the other experiments were run using professional headphones, either in a booth in a perception lab or using a laptop in a quiet room. Intensity was set at a comfortable level for each listener. In all cases the stimuli were presented in a blocked random order, fixed between participants in Experiments 2 and 4a and generated at run time in the others. In all cases, participants heard a randomly-ordered full set of stimulus types in both AB and BA order (i.e. two occurrences of each stimulus type in the matrix for a given experiment, one in each order), followed by the same set in a different random order, and so on until the full set had been presented five times. Participants were given regular opportunities for self-timed breaks; in most of the experiments these opportunities were offered at the end of each full stimulus set.

Experiment 4b deviated somewhat from the summary just given. As noted in Table 1, this experiment was intended to explore hemispheric differences in pitch processing, and stimuli were presented to one ear with white noise in the other ear. Each stimulus was presented four times in each order to each ear, for a total of 16 times, instead of 10 times as in the other experiments. We found no link between monaural presentation and the distribution of F0 and spectral responses, and for purposes of the present report we pool left and right ear presentations for each cell in the stimulus matrix.

Experiments 4a and 4c provide two different measures of test-retest reliability. In Experiment 4a, assessing reliability was a specific goal of the larger study: participants were retested on exactly the same material using exactly the same procedures after an interval of one to two weeks. As for Experiment 4c, it consisted of three separate blocks of stimuli for two separate studies, all run in a single experimental session lasting approximately 45 minutes. The first block contained the stimuli for the basic MF task just sketched and served as a baseline for comparison with the other two blocks. The second block contained the stimuli for a separate study on the 'tritone paradox' (Deutsch 1991, Repp 1994) and is not relevant here

except insofar as it served as a distractor between the first and third blocks. The third block contained a set of stimuli identical in all respects to those of the first block except that they were much shorter (Tone A and Tone B were each 180 ms long, with a gap between the tones of 20 ms). This manipulation was exploratory, to see if stimulus duration would affect listeners' perception; many participants complained that the short-stimulus task was much more difficult, but in the event, duration had little effect on individual patterns of responses. Consequently, we report the results of the first and third blocks of this experiment as a separate measure of reliability.

Data reduction and analysis

The patterns of responses are broadly similar in all seven of our experiments, and except where specified the analyses reported here are for all experiments pooled. The retest data from Experiments 4a and 4c are not included in these pooled analyses. A breakdown of the results by experiment is presented in the online appendix. All analyses were conducted using `R` (R Development Core Team, 2012) and some of its libraries.

We used the Schneider Index (SI), introduced above in Equation 1, to provide a basic quantitative characterisation of each participant's behaviour. Given the structure of the stimulus space in our experiments, for each participant we calculated SI separately for ***each cell*** in the stimulus matrix and then took the average of the individual cell SIs to arrive at a single overall SI for each participant. As noted above, SI ranges from –1.0 to +1.0, with 100% F0 responses yielding a SI of –1.0 and 100% spectral responses yielding a SI of +1.0. As we pointed out in footnote 3, this polarity is opposite to that given in the published formula in Schneider et al (2005).

A substantial minority of participants respond so consistently that there is a SI of either +1 or –1 in almost all cells.[4] By definition, participants with overall SI near +1 or –1 have a pattern of responses that is, in the terms suggested by Seither-Preisler et al., both consistent and homogeneous. These are the participants who can be classified confidently as either F0 or spectral listeners. However, because each participant gave 10 responses in each cell of the stimulus matrix, computing the SI for each participant in each cell of the stimulus matrix allowed us to gain a fairly clear idea of the consistency and homogeneity of the responses of participants whose overall SI lies nearer the middle of the scale. We also discovered, unexpectedly, that the intermediate SI values sometimes reflect the presence of an order of presentation effect: some participants gave different responses depending on whether the two tones of the stimulus were in AB or BA order, and moreover the size of this effect differed in different parts of the stimulus matrix.

To better understand the patterning of the cell-level SI and to investigate the structure of the participants' responses, we also carried out a Principal Components Analysis

[4] In fact, in Experiment 1, we were able to detect a faulty stimulus because, in one cell of the matrix, about a third of the participants had SI near 0 despite having SI near +1 or –1 in all the other cells. Investigation revealed that Tone A in the AB-order stimulus in that cell had incorrect partials, such that both F0 and spectral listening should yield the same response. This anecdote gives an idea of the consistency and reliability with which some participants respond. On the other hand, the fact that some participants respond very consistently should not blind us to the fact that many others do not.

(Jolliffe, 2002) for each experiment separately. We found that the first two principal components are extremely similar across experiments: the first principal component, which explains at least half of the variance, is equivalent to the overall SI, while the second component expresses one of the clear response patterns at intermediate SI values. We converted this second principal component to an index that ranges, like SI, from –1 to +1; we refer to this measure as the Consistency Index (CI). Further detail on our mathematical treatment of the Principal Components data is given in the online appendix; further detail on the interpretation of CI is given in the next section.

Results

Individual differences between participants

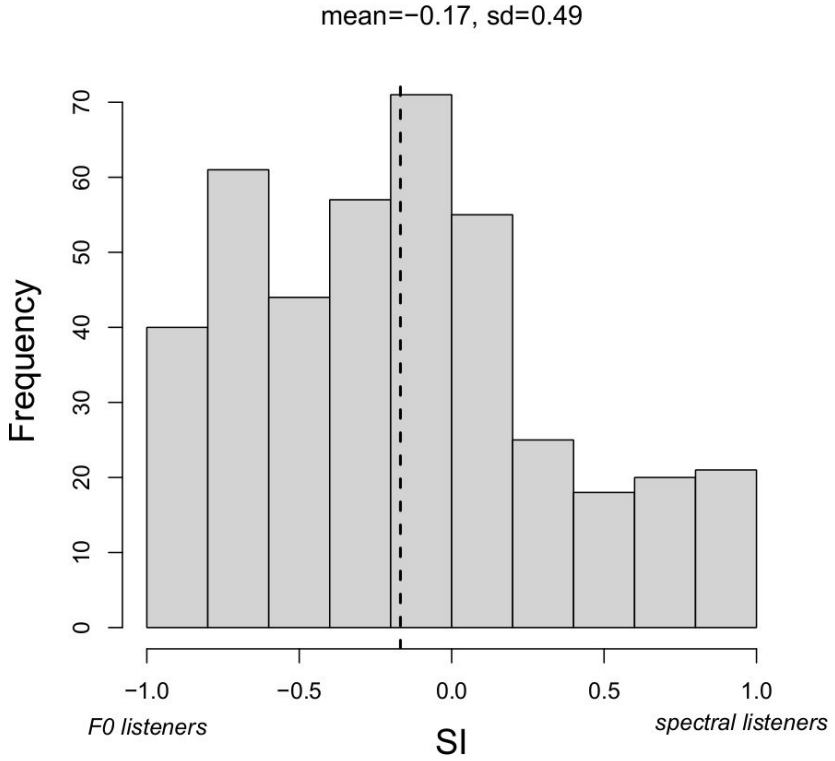

Figure 2. Distribution of Schneider Index (-1.0 = F0 listener, +1.0 = spectral listener) for all 412 participants in the seven experiments (retest data from Experiments 4a and 4c excluded); dashed vertical line represents the mean. This figure may be compared to Fig. 1b in Schneider et al. 2005.

Basic distribution of SI. Our findings replicate those of the studies already discussed, in the sense that we find a range of response patterns from SI near –1.0 to SI near +1.0. Figure 2 shows the distribution of SI for all experiments pooled. The distribution does not appear to be Gaussian (Q-Q plot and Shapiro-Wilk normality test $W = 0.964$, $p < .0001$), but contrary to what is reported especially by Schneider et al., it is not bimodal either (Hartigans' dip test for unimodality $D = 0.018$, $p = 0.48$; Hartigan & Hartigan, 1985). There appears to be a bias toward F0 responses, as suggested by Seither-Preisler et al. There is also a single clear mode about 0, which seems to correspond to the substantial number of participants that Seither-Preisler et al. excluded for 'guessing'. The difference between our findings and Schneider et

al.'s is apparently attributable to the controls for octave-shifted percepts that Schneider et al. included in their study (see above). Schneider (personal communication, August 2012) informs us that if octave-shifted percepts are counted as F0 percepts, then the overall distribution of SI in their data looks more like what we report here: a greater proportion of F0 listeners, and a more Gaussian shape**Principal Components Analysis.** Principal Components Analysis, carried out on each experiment separately, confirms the validity of the overall SI as the principal expression of the individual differences on the MF task. In every experiment, the first principal component, explaining between 46.5% and 80.3% (mean 65.7%) of the between-participants variance, is effectively equivalent to SI. As noted above, a common second principal component also emerged for all experiments, explaining between 6.5% and 18.4% (mean 12.1%) of the variance, with eigenvalues ranging between 0.47 and 3.24 (mean 1.05), and correlating only weakly with the first component ($r = -.10$, $p < .05$). Principal components other than the first two were inconsistent across experiments, had eigenvalues less than 0.5 except in Experiments 3a and 3b, and were not obviously interpretable. Further detail on the Principal Components Analysis is given in the online appendix.

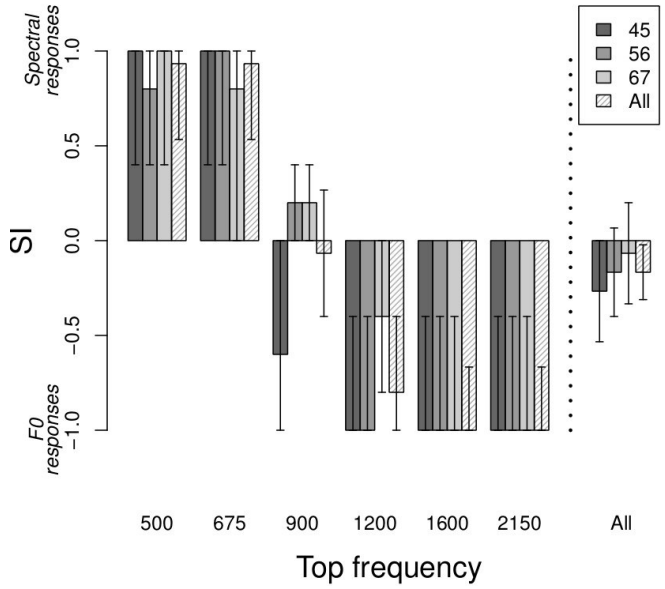

Figure 3. Individual responses of one 'inhomogeneous' participant from Experiment 4c, representative of roughly 7.5% of participants who give mostly spectral responses to stimuli with low frequency level and mostly F0 responses to those with high frequency level.

Inspection of the data showed that the second principal component reflects a consistent difference between responses in the lower and higher portions of the top frequency dimension of the stimulus matrix. Closer investigation of the data from participants with SI values in the middle of the scale showed that some give predominantly spectral responses at lower frequency levels and predominantly F0 responses at higher levels. Figure 3, which shows the responses of one such participant, illustrates this pattern graphically. This specific pattern of responses seems to be what is captured by the Consistency Index (CI). Importantly, CI is skewed towards positive values (minimum -0.53, maximum 0.96, median 0.09, mean

0.13). Only two participants of the entire group of 412 had CI values as low as ~ -0.5; removing them increased the minimum to –0.28. This suggests that one theoretically possible pattern of responses, namely F0 responses to stimuli with lower overall frequency and spectral responses to those with higher overall frequency, occurs only very rarely, while the opposite pattern is quite common.

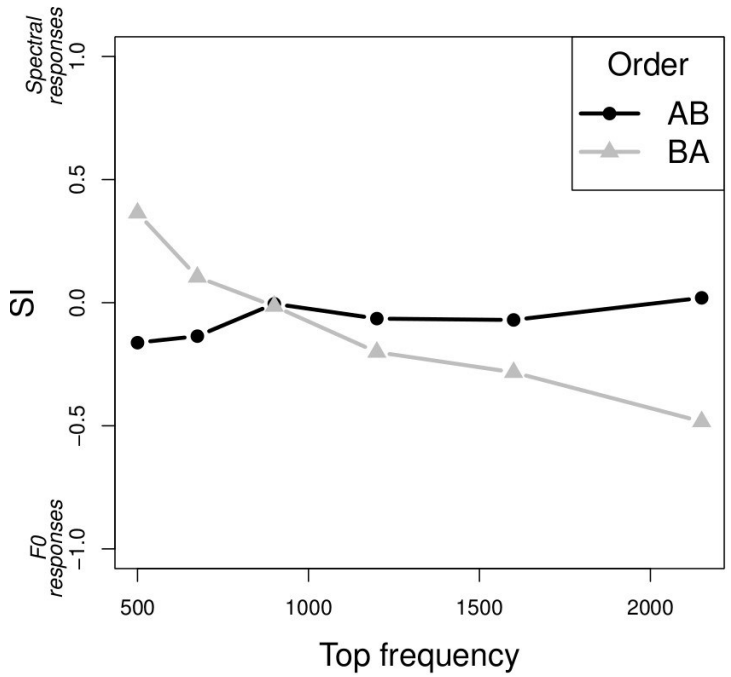

Figure 4. Interaction between order of presentation and frequency level, as shown by pooled data from the Experiment 4 set. Data from the other experiments (which have different specific frequency levels) are qualitatively very similar. It can be seen clearly that BA stimuli show a tendency to elicit more F0 responses at higher frequency levels, while AB stimuli do not. This interaction is highly significant (on an ANOVA, $F(5,2148) = 3.3 \times 10^{29}$, $p < .001$).

Order of presentation effect. In addition to the pattern expressed by CI, we found to our surprise that a number of participants show an order of presentation effect that interacts with frequency level. Specifically, at low frequency levels, AB stimuli (where the missing fundamental falls) are more likely than BA stimuli (where the missing fundamental rises) to elicit F0 responses, while at high frequency levels the reverse is true. An alternative way of stating this observation, which may ultimately provide more insight into its cause, is to say that low frequency level favours 'down' responses and high frequency level favours 'up' responses. Inspection revealed that this tendency does not affect all participants equally: some show no influence of order of presentation at all, while others show dramatic differences between low and high frequency level. Overall, however, the pooled data reflect this tendency, as can be seen in Figure 4; as can also be seen, frequency level affects the response to BA stimuli more than the response to AB stimuli.

For the purpose of further analyses, we quantified this order effect by taking the mean of the absolute value of the difference in SI between AB and BA responses at each

frequency level. There are weak but significant correlations between the order effect, so quantified, and both SI and CI (for SI, $r = .19$, $p < .001$ (Spearman's $\rho = .30$, $p < .001$); for CI, $r = -.20$, $p < .001$).

Test-retest reliability. As explained in the 'Procedure' section, we report two different assessments of the test-retest reliability, not only for SI, but also for CI and the order effect. In Experiment 4a (a conventional test of reliability involving exact repetition of the experiment after an interval of one to two weeks), the correlation between test and retest for SI was $r = .87$, $p < .001$.[5] CI also showed high test-retest reliability ($r = 0.83$, $p < .001$); the order effect slightly less so ($r = 0.52$, $p < .001$). In Experiment 4c, the second test – presented later in the same experimental session, as described in the 'Procedure' section – involved stimuli that differed in duration but were otherwise identical to those of the first test. Here the correlation between test and retest was $r = .94$, $p < .001$ for SI, $r = 0.73$, $p < .001$ for CI, and $r = 0.74$, $p < .001$ for the order effect. It therefore seems clear that the individual differences tapped by the MF task are very robust.

Effect of stimulus variables

Overall frequency level and spectral composition. We saw in the introduction that both Schneider et al. and Seither-Preisler et al. report more F0 responses to stimuli with overall higher frequency, though it is not easy to tell whether the effect is primarily due to higher frequency as such or to higher harmonic rank of the partials (spectral composition, in our terms). Our experiments strongly suggest that actual frequency is the more important factor, though the order effect illustrated in Figure 4 may be relevant too. The upper panel of Figure 5 shows the effect of frequency level on the overall distribution of responses for all experiments pooled. (Frequency levels above 2200, which were used only in Experiments 3a and 3b and which yielded an overwhelming preponderance of F0 responses, are excluded.) It can be seen that at very low frequency levels (500 Hz and below, implying MF values between 30 and 100 Hz) the distribution of SI is nearly Gaussian, with a mode near 0, i.e. showing no preference for either spectral or F0 responses. The distributions then become flatter, with considerable numbers of clear F0 listeners and spectral listeners in the mid-range of frequencies, around 1000 Hz. As the frequency level increases from the mid-range, the distribution becomes increasingly skewed toward F0 responses.

For comparison, the lower panel of Figure 5 presents a similar analysis, showing the effect of spectral composition on patterns of responses. (It is based only on the Experiment 4 set, which shared the same spectral composition variables.) It is difficult to interpret as clearly as the left-hand panel, and there is certainly no obvious trend. As discussed in the introduction, there is a significant degree of interdependence between the stimulus variables, and our data do not permit us to explore this issue further.

[5] This may be compared to Schneider et al.'s report (2005: 1046) of test-retest reliability of $r = .96$ for a subgroup of 37 participants retested six months after the original experiment.

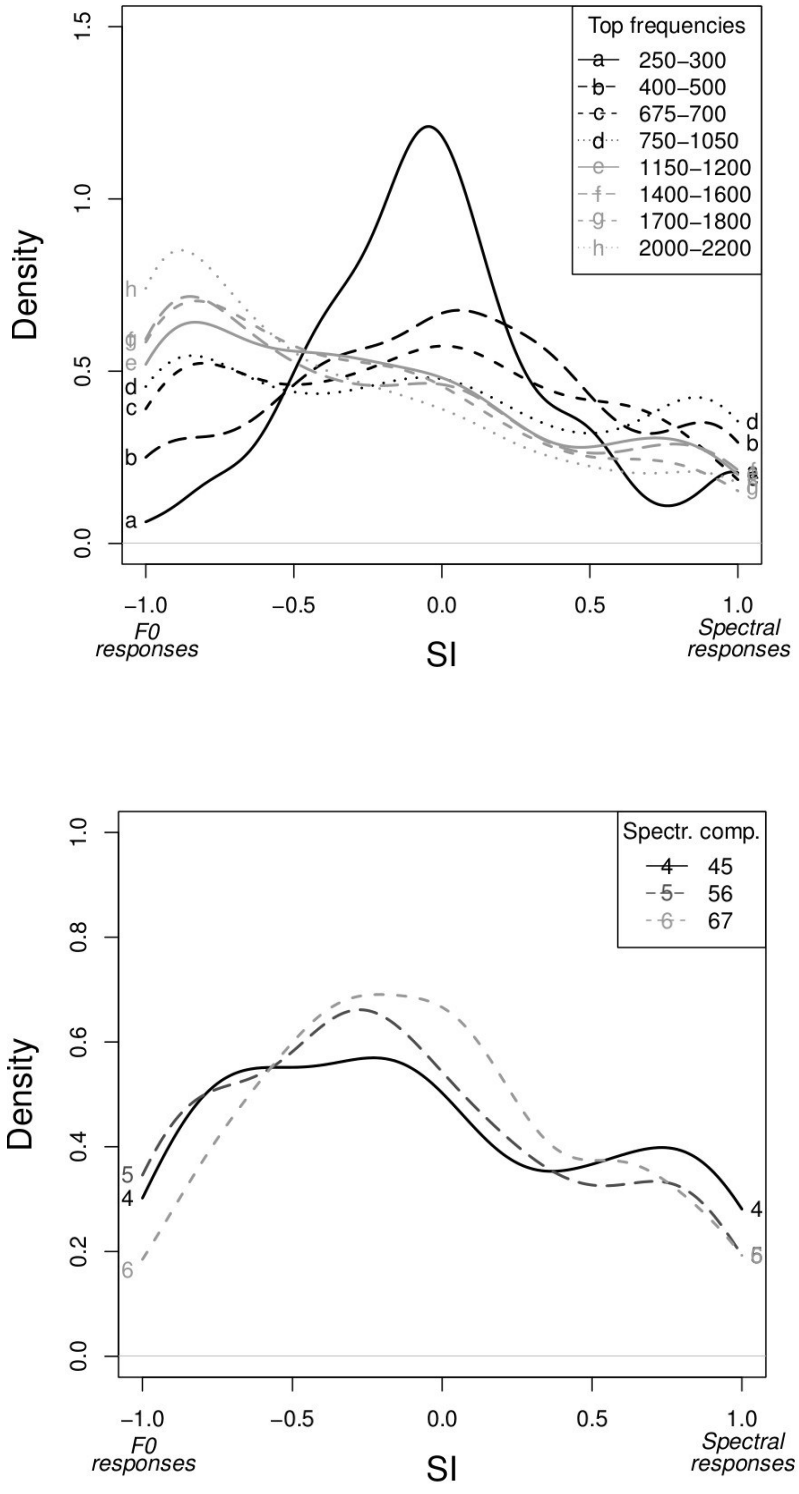

Figure 5. Effects of overall frequency level and spectral composition on patterns of responses. The lines represent envelope-like approximations of the distributions, computed by Kernel Density Estimation (Silverman, 1986). Panel A (upper panel) shows clearly that higher frequency levels give rise to more F0 responses; this panel may be compared to Fig. 1c in Schneider et al. 2005. Panel B (lower panel) is more difficult to interpret. For further discussion see text.

Cluster analysis

Given the apparent diversity of response patterns, we subjected the data to a *k*-means cluster analysis, locating every participant in a three dimensional space defined by SI, CI, and the order effect. We found that seven clusters fit across all experiments. (Technical details are given in the online appendix). Figure 6 shows these clusters plotted in two different two-dimensional projections, one showing the relation between SI and CI, and one showing the relation between SI and the order effect. The clusters are clearly interpretable. Cluster 4 (black triangles, ▲) comprises consistent spectral listeners and Cluster 5 (grey crosses, +) comprises consistent F0 listeners. Cluster 6 (grey rectangles, □) comprises weak spectral listeners, while Clusters 7 (black circles, ●) and 1 (grey inverted triangles, ▽) comprise weak F0 listeners, some of whom show clear effects of frequency level on their pattern of responses. Cluster 2 (black x's, ×) comprises listeners with no clear preference and/or those strongly affected by the order effect; Cluster 3 (grey diamonds, ♦) comprises listeners who show a strong shift of listening preference from lower to higher frequency levels.

Note that roughly a quarter (22.8%) of all participants fall into Cluster 2 (listeners with no clear preference), about the same proportion excluded as inconsistent by Seither-Preisler et al. However, it can be seen from Figure 6 that this cluster is the one most strongly affected by the order effect. That is, for many of these participants the fact that SI is near zero results from a consistent pattern of responses – but one which cannot be expressed by the SI and which, as we suggested earlier, may actually undermine the assumptions underlying the MF task. Further research is clearly called for.

Effect of participant variables

Recall that both Schneider et al. and Seither-Preisler et al. were interested in the effects of musical training. Unlike either of those studies, we had very few professional musicians among our participants, but we did collect self-reports on musical activity and training, and on this basis we can very roughly classify our participants as either musical (generally corresponding to Seither-Preisler et al.'s amateur musicians) or not (Seither-Preisler et al.'s non-musicians). Like Schneider et al., we find no effect of musicality on SI ($t(388.48) = -1.27$, $p=0.20$), but there are significant effects on the order effect ($t(424.12) = 5.99$, $p < .001$), with musical participants being significantly less susceptible to order effects than non-musical participants. Musical participants were also slightly less influenced by frequency level, as expressed by CI ($t(361.48) = -4.49$, $p < .001$). This is at least consistent with the idea that musical listeners are performing the task as intended, i.e. hearing Tone A and Tone B separately and judging their relative pitch level, while non-musical listeners may be treating the pair of tones as some sort of holistic unit.

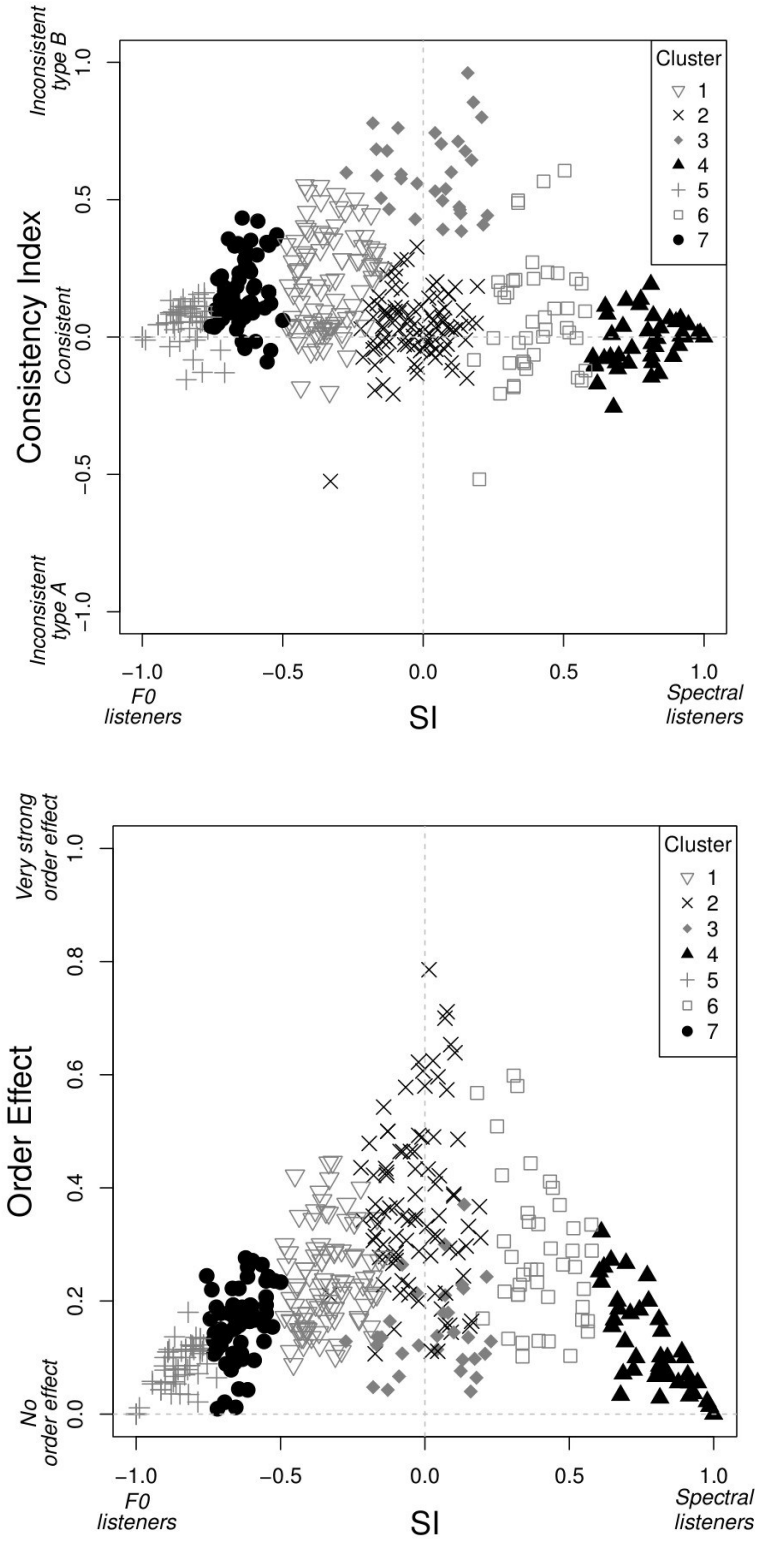

Figure 6. Scatterplots of the 7 k-means clusters for our data. Only two 2-dimensional projections are shown here: Panel A plots the Schneider Index against the Consistency Index, while Panel B shows the Schneider Index plotted against the order effect. For further discussion see text.

Schneider and Wengenroth (2009) found no effect of age or gender on SI. We find no effect of gender, but a slight effect of age, with older participants slightly more likely to give spectral responses. On the full set of 412 participants, the correlation between age and SI is $r = .16$, $p < .01$; to compensate for the very skewed age distribution in our participant group, we ran the same analysis based only on participants over 25, and found the same tendency but with too little statistical power to reach significance ($r = .20$, $.05 < p < .10$). One can imagine a variety of explanations for an age effect on SI; probably the most plausible is one based on physical changes in the inner ear, but cortical changes can by no means be ruled out (cf. Whitfield, 1980). We find no effect of age or gender on either CI or the susceptibility to the order effect.

Discussion and Conclusion

Our investigations have confirmed that there are robust individual differences in the perception of MF stimuli. As we pointed out in the introduction, a comparison of two recent studies (Schneider et al., 2005; and Seither-Preisler et al., 2007) suggests that these differences will emerge even from experiments that substantially diverge methodologically. In the present set of studies we have demonstrated specifically that individuals' responses are unaffected by large differences in stimulus duration, and our test-retest reliability results confirm that individuals' responses are consistent over time. At the same time, we have shown that it is something of an oversimplification to classify individuals as 'F0 listeners' or 'spectral listeners'.

First, we have shown that as many as a quarter of all individuals appear to have no consistent response preference at all. Superficially, this finding diverges from the results of Schneider et al. (2005) (who report a strongly bimodal U-shaped distribution that appears to justify a binary classification of participants), but as noted earlier the difference may be explainable by Schneider et al.'s careful control of octave-shifted percepts. Our findings more obviously agree with those of Seither-Preisler et al. (2007), who excluded roughly a quarter of their participants from further analysis on the grounds that they were guessing. In separate analyses associated with Experiment 4c, we found that some of these inconsistent listeners also respond inconsistently to stimuli used to investigate the 'tritone paradox' (Deutsch 1991, Repp 1994), which might suggest that their responses reflect a more general difficulty with judging pitch or pitch direction, or alternatively, perhaps, a susceptibility to octave-shifted percepts. In any case, the conclusion that such individuals are merely 'guessing' seems decidedly premature, because of the fact, illustrated in Figure 6, that many of them actually do show a consistent response pattern that simply happens not to be captured by data reduction in terms of SI. This group of listeners needs to be treated separately in drawing conclusions about MF perception, and may be interesting to study in its own right.

Second, we have confirmed and extended others' findings that certain stimulus variables have predictable effects on responses to MF stimuli. The effect of overall frequency level is strong enough that 7.5% of individuals (in our cluster analysis, those in cluster 3) give consistently opposite responses in different areas of the stimulus space, responding as 'spectral listeners' at low overall frequencies and as 'F0 listeners' at high overall frequencies. Studying such listeners may provide useful insight into the sources of the two different modes of perceiving MF stimuli. Even

among participants who are not so strikingly affected, we have shown that in general, responses are influenced by the overall frequency level and perhaps by the spectral composition of the stimuli. This has implications for the construction of appropriate stimuli in further research.

Our findings on stimulus and participant variables in MF perception should make it possible for researchers whose interest is in the physical and psychophysical foundations of the phenomenon to make more confident methodological choices, and may help shed light on apparent discrepancies in the results of different studies. It should now also be possible to use the MF task with greater methodological confidence in studies that are not essentially concerned with the phenomenon itself, but with what it tells us about individual differences more generally. For example, it could be revealing to determine the heritability of modes of MF perception, or to investigate the relationship between MF perception and other perceptual and cognitive tasks, from basic auditory sensitivity to language-related tasks such as non-word repetition and digit span. It may also be interesting to investigate brain structure and function (as in the studies by Patel & Balaban, 2001, or Schneider et al., 2005) with a more fine-grained characterisation of individual behavioural differences than simply ‘F0 listener’ and ‘spectral listener’. We believe we have provided the research community with a better-calibrated tool for all these purposes.

References

Bernstein, J.G., & Oxenham, A.J. (2006). The relationship between frequency selectivity and pitch discrimination: effects of stimulus level. *Journal of the Acoustical Society of America,* 120, 3916-3928.

Caldwell-Harris, C., Biller, A., Ladd, D. R., Dediu, D., & Christiansen, M. (2012). Musical ability and prior tone language experience facilitate learning an artificial tone language. Paper presented at the American Association for Applied Linguistics, Boston, March 2012.

Deutsch, D. (1991). The tritone paradox: An influence of language on music perception. *Music Perception,* 8, 335-347.

Faulkner, A. (1985). Pitch discrimination of harmonic complex signals: Residue pitch or multiple component discrimination? *Journal of the Acoustical Society of America*, 78, 1993-2004.

Gockel, H. E., Plack, C. J., & Carlyon, R. P. (2005). Reduced contribution of a nonsimultaneous mistuned harmonic to residue pitch. *Journal of the Acoustical Society of America,* 118, 3783-3793.

Gockel, H. E., Carlyon, R. P., & Plack, C. J. (2010). Combining information across frequency regions in fundamental frequency discrimination. *Journal of the Acoustical Society of America,* 127, 2466-2478.

Hartigan, J. A., & Hartigan, P. M. (1985). The dip test of unimodality. *Annals of Statistics,* 13, 70–84.

Jolliffe, I. (2002). *Principal Component Analysis*. New York: Springer Verlag.

Licklider, J. C. R. (1951). A duplex theory of pitch perception. *Experientia*, 7, 128-134.

Meddis, R., & Hewitt, M. J. (1991). Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification. *Journal of the Acoustical Society of America*, 89, 2866-2882.

Moore, B. C. J. (2012). *An Introduction to the Psychology of Hearing,* sixth edition. Bingley UK: Emerald Group Publishing.

Moore, B. C. J., Glasberg, B. R., & Peters, R. W. (1985). Relative dominance of individual partials in determining the pitch of complex tones. *Journal of the Acoustical Society of America*, 77, 1853-1860.

Moore, B. C., J., & Gockel, H. E. (2011). Resolvability of components in complex tones and implications for theories of pitch perception. *Hearing Research,* 276, 88-97.

Patel, A., & Balaban, E. (2001). Human pitch perception is reflected in the timing of stimulus-related cortical activity. *Nature Neuroscience,* 4, 839-844.

Plomp, R. (1967). Pitch of complex tones. *Journal of the Acoustical Society of America*, 41, 1526-1533.

R Development Core Team (2012). *R: A language and environment for statistical computing. R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Repp, B. H. (1994). The tritone paradox and the pitch range of the speaking voice: A dubious connection. *Music Perception,* 12, 227-255.

Ritsma, R. J. (1962). Existence region of the tonal residue. I. *Journal of the Acoustical Society of America*, 34, 1224-1229.

Ritsma, R. J. (1963a). Existence region of the tonal residue. II. *Journal of the Acoustical Society of America*, 35, 1241-1245.

Ritsma, R. J. (1963b) On pitch discrimination of residue tones. *International Audiology*, 2, 34-37.

Schneider, P., Sluming, V., Roberts, N., Scherg, M., Goebel, R., Specht, H. J., … Rupp, A. (2005). Structural and functional asymmetry of lateral Heschl's gyrus reflects pitch perception preference. *Nature Neuroscience,* 8, 1241-1247.

Schneider, P., & Wengenroth, M. (2009). The Neural Basis of Individual Holistic and Spectral Sound Perception. *Contemporary Music Review*, 28, 315-328.

Schouten, J. F. (1940). The residue and the mechanism of hearing. *Proceedings of the Koninklijke Akademie van Wetenschap*, 41, 1086-1093.

Seither-Preisler, A., Johnson, L., Krumbholz, K., Nobbe, A., Patterson, R., Seither, S., & Lütkenhöner, B. (2007). Tone sequences with conflicting fundamental pitch and timbre changes are heard differently by musicians and nonmusicians. *Journal of Experimental Psychology: Human perception and performance,* 33, 743-751.

Sidtis, J. J. (1980) On the nature of cortical function underlying right hemisphere auditory functions. *Neuropsychologia*, 18, 321-330.

Silverman, B. W. (1986). *Density Estimation.* London: Chapman and Hall.

Smoorenburg, G. (1970). Pitch perception of two-frequency stimuli. *Journal of the Acoustical Society of America,* 48, 924-942.

Terhardt, E. (1974). Pitch, consonance, and harmony. *Journal of the Acoustical Society of America,* 55, 1061-1069.

Terhardt, E. (1979). Calculating virtual pitch. *Hearing Research*, 1, 155-182.

Whitfield, I. C. (1980). Auditory cortex and the pitch of complex tones. *Journal of the Acoustical Society of America*, 67, 644-647.

Appendix

Online Supplementary Materials for
Patterns of individual differences in the perception of
missing-fundamental tones

D. Robert Ladd[1], Rory Turnbull[1], Charlotte Browne[1], Catherine Caldwell-Harris[2], Lesya Ganushchak[3], Kate Swoboda[1], Verity Woodfield[1], and Dan Dediu[3]

[1]School of Philosophy, Psychology and Language Sciences, University of Edinburgh
[2]Department of Psychology, Boston University
[3]Max-Planck-Institute for Psycholinguistics, Nijmegen

Corresponding author: Prof. D. Robert Ladd, School of Philosophy, Psychology and Language Sciences, University of Edinburgh, 3 Charles Street, Edinburgh EH8 9AD, Scotland. Email bob.ladd@ed.ac.uk

Summary

This appendix forms part of the Supplementary Online Materials and contains more details about the stimuli and the statistical analyses and results described in the main paper.

The Stimuli

Sound files of selected stimuli are provided here, arranged so that readers can easily make some of the comparisons discussed in the main paper. Except where indicated, the stimuli illustrated here have phase-controlled partials (see 'Stimulus preparation' in the main paper).

The stimuli are identified here by code numbers consisting of three segments, e.g. 56.1600.BA. The first segment refers in abbreviated form to the spectral composition; the second gives the top frequency in Hz; and the third specifies the order of the two component tones (AB or BA). The code for spectral composition indicates the highest harmonics in Tone A and Tone B: thus '56' refers to stimuli in which Tone A consists of harmonics 3, 4 and 5 and Tone B consists of harmonics 4, 5 and 6. For more detail see Table 1 in the main paper.

'Up' responses indicate a spectral percept with AB stimuli and an F0 percept with BA stimuli. 'Down' responses indicate a spectral percept with BA stimuli and an F0 percept with AB stimuli.

Order. Here are two pairs of stimuli that are identical except for order:

45.0675.AB 45.0675.BA
67.1600.AB 67.1600.BA

Spectral composition. Here are two triplets of stimuli that are identical except for spectral composition:

45.1200.AB 56.1200.AB 67.1200.AB
45.2150.BA 56.2150.BA 67.2150.BA

These are 'identical' only in the context of our stimulus matrix, i.e. in the sense that the top frequency and the order are the same. The missing fundamentals, and the intervals between them, are different in each case, because of the interdependencies discussed in the section 'Stimulus variables in the MF task' in the main paper.

Top frequency. Here is a set of stimuli that are identical except for top frequency:

45.0500.AB 45.0675.AB 45.0900.AB 45.1200.AB 45.1600.AB 45.2150.AB

Readers may observe that their response changes from 'up' to 'down' as the top frequency increases. This is the pattern of responses illustrated in Figure 3 of the main paper and may be related to the order effect illustrated in Figure 5 of the main paper.

Phase. Here is a pair of stimuli that differ only in whether phase was controlled in generating them:

56.0500.BA (phase controlled)
56.0500.BA (phase not controlled)

Duration. Here are two pairs (from the test-retest material in Experiment 4c) that differ only in duration:

67.0900.AB (500 ms tones, 250 ms gap) 67.0900.AB (180 ms tones, 20 ms gap)
45.0500.BA (500 ms tones, 250 ms gap) 45.0500.BA (180 ms tones, 20 ms gap)

The Schneider Index

As described in the main paper, each stimulus is described by a *Top Frequency Level* (here denoted *FL*), a *Spectral Composition* (denoted *SC*), and a *Direction* (AB or BA) and any single participant gives an "Up" or "Down" response for each such stimulus. By definition, we created the stimuli in such a way that a "Down" response to the AB order and an "Up" response to the BA order are based on the missing fundamental (*F0* responses), while "Up" for AB and "Down" for BA are based on the harmonics present in the stimuli (*spectral* responses). See Figure 1 and associated discussion in the main paper for more details.

For a set of stimuli and responses, we can count the number of F0 (denoted f0) and spectral (denoted sp) responses; with these, the *Schneider Index* (denoted SI) for this set of stimuli is:

$$\mathrm{SI} = \frac{\mathrm{sp} - \mathrm{f0}}{\mathrm{sp} + \mathrm{f0}} \tag{1}$$

which can vary between –1 (100% F0 responses) and +1 (100% spectral responses). When considering *all* the responses given by the participant in all FL and SC conditions and both orders, we compute the *overall Schneider Index*, denoted here as SI_{O}, but we can also compute *partial Schneider Indices* by restricting the value of some experimental parameters, such as for a given FL value (say, 500; denoted $\mathrm{SI}_{\mathrm{FL=500}}$), a given SC value (say, using the notations defined above, 45; $\mathrm{SI}_{\mathrm{SC=45}}$), or for a single cell in the FL×SC stimulus matrix (say FL=500 and SC=45; $\mathrm{SI}_{\mathrm{FL=500,SC=45}}$), or even for a given order (such as $\mathrm{SI}_{\mathrm{AB}}$). While measures comparable to the overall SI_{O} have been used in the previous literature as a measure of a participant's global style, the various partial SI allow us a better understanding of the individual differences in the perception of the missing fundamental and the factors that affect them.

Summaries based on the partial Schneider Indices

Using various partial SI measures, we can define a number of summaries capturing different aspects of the participants' behaviour. One such summary captures the magnitude of the *order effect*, quantifying the difference that the two orders of presentation (AB and BA) might have on the participants' answers. We will denote this as *OE*, defined as the mean absolute difference between the partial SI for the AB and BA orders:

$$\mathrm{OE} = \underset{fl,sc}{mean} \frac{|\,\mathrm{SI}_{\mathrm{FL}=fl,\mathrm{SC}=sc,\mathrm{AB}} - \mathrm{SI}_{\mathrm{FL}=fl,\mathrm{SC}=sc,\mathrm{BA}}\,|}{2} \tag{2}$$

where *fl* is a valid FL value, *sc* is a valid SC value, $|x|$ is the absolute value (modulus) of x, and division by 2 insures that OE lies between 0 (no order effect), and 1 (maximum possible order effect, i.e. consistently opposite responses to AB and BA items).

Other summaries are driven by the structure of the data, in the sense that they are derived from the principal components in a *Principal Components Analysis* (*PCA*; Jolliffe 2002). PCA is a technique which transforms a set of N inter-correlated variables into the same number of independent components ordered by the amount of variation in the data they explain. Thus, the first component, PC_1, explains most of the variation in the data, followed by PC_2, and so on. Given that these components are linear combinations of the original variables weighted by their loadings, these loadings can be used to interpret the meaning of the components. We performed PCA on the cell-level SI as the N inter-correlated variables: $\mathrm{SI}_{\mathrm{FL}=fl,\mathrm{SC}=sc}$, where *fl* and *sc* represent valid values of FL and SC as above.

As described in the main paper, we found that the first and second components, PC_1 and PC_2, are extremely similar across experiments in both structure (i.e., the loadings of the cell-level SI) and the amount of variance explained (Table S1). More precisely, PC_1 explains about half of the variance and is equivalent to the overall SI (SI_{O}) in the sense that all cell-level SI have loadings of same sign and

comparable magnitude. Therefore, we abstracted away from the actual loadings and defined PC_1^* as

$$PC_1^* = \underset{fl,sc}{mean}\left(SI_{FL=fl,SC=sc}\right) \quad (3)$$

whereby we assign the same weight to all FL×SC cells. As expected, PC_1^* is highly correlated with SI_O (overall SI) ($r = .99$, $p < .001$). In the main paper we therefore used SI_O (there denoted simply SI) rather than PC_1^*.

PC_2 expresses one of the response patterns we identified for participants with intermediate SI_O values, namely a difference between responses at lower and higher FL. Specifically, the loadings of the cell-level SI of the cells with FL lower than a threshold, *T*, have the opposite sign to those of the cells with FL higher than *T*. This is the pattern of responses illustrated in Figure 3 of the main paper. We interpreted this pattern as reflecting a type of consistency in the participant's responses and, as above, we abstracted away from the actual loadings by formalizing it as the *Consistency Index* (denoted *CI*) as follows:

$$CI = \frac{\underset{fl<T,sc}{mean}\left(SI_{FL=fl,SC=sc}\right) - \underset{fl\geq T,sc}{mean}\left(SI_{FL=fl,SC=sc}\right)}{2} \quad (4)$$

In practice, the threshold *T* depends on the actual experiment (its approximate values are 1000Hz, 1200Hz, 1600Hz, 1800Hz, 1000Hz, 1000Hz and 1000Hz for experiments 1, 2, 3a, 3b, 4a, 4b, and 4c, respectively), but we were able to find a general formula relating it to the limits of the FL space. We did this by regressing the observed *T* values on the minimum and maximum FL values across the experiments, as we found that:

$$T^* = 800 + \frac{FL_{max} - FL_{min}}{5} \quad (5)$$

gives a good approximation for the observed threshold values, *T*. CI expresses the size of the difference between low and high FL, with strongly polarised values (near –1.0 and +1.0) representing a large difference, and values near 0.0 representing a small difference. Strongly positive values indicate a switch from spectral to F0 responses as FL increases; strongly negative values would mean the reverse, but as noted in the main paper strongly negative values are never found. Among participants with strongly positive CI, the location of the crossover point (the threshold T^*) on the FL dimension is fairly consistent, somewhere in the vicinity of 1000 Hz.

Table S1
Principal Components Analysis (PCA) for each experiment.

Experiment / **Cell (SC, FL)**	**PC_1**	**PC_2**
Experiment 1		
Eigenvalues: **PC_1 (3.79)**, **PC_2 (1.16)**, PC_3 (0.45)		
Variance explained	***56.99%***	***17.39%***
56, 300	*-0.27*	*-0.25*
56, 500	*-0.28*	*-0.29*
56, 900	*-0.30*	0.02
56, 1400	*-0.22*	**0.34**
56, 2200	*-0.17*	**0.19**
78, 300	*-0.24*	*-0.32*

78, 500	*-0.31*	*-0.32*
78, 900	*-0.34*	0.02
78, 1400	*-0.26*	**0.36**
78, 2200	*-0.17*	**0.24**
90, 300	*-0.13*	*-0.22*
90, 500	*-0.33*	*-0.23*
90, 900	*-0.31*	0.09
90, 1400	*-0.24*	**0.37**
90, 2200	*-0.19*	**0.28**
Experiment 2		
Eigenvalues: **PC_1 (3.55)**, PC_2 (0.47), PC_3 (0.14)		
Variance explained	***80.31%***	***10.67%***
56, 500	*-0.30*	**0.47**
56, 750	*-0.37*	**0.21**
56, 1050	*-0.34*	**0.10**
56, 1400	*-0.35*	-0.04
56, 1800	*-0.29*	*-0.36*
89, 500	*-0.32*	**0.49**
89, 750	0.00	0.05
89, 1050	*-0.35*	-0.05
89, 1400	*-0.34*	*-0.41*
89, 1800	*-0.32*	*-0.43*
Experiment 3a		
Eigenvalues: **PC_1 (10.02)**, **PC_2 (3.24)**, **PC_3 (1.06)**		
Variance explained	***56.84%***	***18.36%***
56, 250	-0.03	-0.08
56, 500	*-0.12*	*-0.22*
56, 750	**0.18**	*-0.31*
56, 1050	*-0.21*	*-0.25*
56, 1400	*-0.20*	*-0.22*
56, 1800	*-0.20*	0.02
56, 2000	*-0.24*	0.01
56, 2500	*-0.21*	**0.19**
56, 3000	*-0.19*	**0.12**
56, 4000	*-0.24*	**0.22**
56, 5000	*-0.23*	**0.26**
56, 6000	*-0.13*	**0.31**
78, 300	0.02	*-0.15*
78, 500	-0.07	*-0.18*
78, 900	*-0.22*	*-0.28*
78, 1400	*-0.22*	*-0.15*
78, 2200	*-0.20*	**0.12**
89, 250	0.01	*-0.11*
89, 500	-0.06	*-0.15*
89, 750	*-0.14*	*-0.30*
89, 1050	*-0.23*	*-0.22*
89, 1400	*-0.23*	-0.06
89, 1800	*-0.25*	0.03
89, 2000	*-0.24*	0.07
89, 2500	*-0.23*	**0.11**
89, 3000	*-0.24*	**0.11**
89, 4000	*-0.19*	**0.18**
89, 5000	*-0.17*	**0.23**
89, 6000	-0.06	**0.10**
Experiment 3b		
Eigenvalues: **PC_1 (3.90)**, **PC_2 (1.13)**, PC_3 (0.55)		
Variance explained	***46.50%***	***13.46%***
56, 250	**0.12**	-0.09

56, 300	**0.18**	-0.04
56, 400	**0.20**	**0.13**
56, 500	**0.24**	0.07
56, 700	**0.26**	**0.24**
56, 900	**0.26**	**0.13**
56, 1150	**0.24**	0.09
56, 1400	**0.24**	0.06
56, 1700	**0.22**	0.00
56, 2000	**0.19**	*-0.20*
56, 2500	**0.16**	*-0.25*
56, 3000	**0.14**	*-0.27*
56, 4000	**0.13**	*-0.31*
56, 5000	**0.13**	*-0.34*
89, 250	-0.01	0.06
89, 300	0.05	-0.02
89, 400	**0.15**	**0.14**
89, 500	**0.18**	**0.20**
89, 700	**0.24**	**0.23**
89, 900	**0.23**	**0.19**
89, 1150	**0.25**	**0.19**
89, 1400	**0.24**	0.05
89, 1700	**0.24**	-0.03
89, 2000	**0.19**	*-0.19*
89, 2500	**0.14**	*-0.20*
89, 3000	**0.14**	*-0.31*
89, 4000	**0.11**	*-0.29*
89, 5000	**0.11**	*-0.23*
Experiment 4a		
Eigenvalues: **PC_1 (3.37)**, PC_2 (0.67), PC_3 (0.26)		
Variance explained	***61.29%***	***12.10%***
45, 500	**0.16**	**0.42**
45, 675	**0.20**	**0.37**
45, 900	**0.31**	0.07
45, 1200	**0.27**	*-0.11*
45, 1600	**0.30**	*-0.23*
45, 2150	**0.29**	*-0.22*
56, 500	**0.17**	**0.33**
56, 675	**0.21**	**0.20**
56, 900	**0.28**	0.08
56, 1200	**0.23**	*-0.20*
56, 1600	**0.23**	*-0.16*
56, 2150	**0.21**	*-0.21*
67, 500	**0.12**	**0.43**
67, 675	**0.16**	**0.26**
67, 900	**0.24**	-0.01
67, 1200	**0.22**	-0.09
67, 1600	**0.27**	*-0.17*
67, 2150	**0.27**	*-0.08*
Experiment 4a (retest)		
Eigenvalues: **PC_1 (3.64)**, PC_2 (0.98), PC_3 (0.33)		
Variance explained	***58.60%***	***15.83%***
45, 500	**0.17**	**0.40**
45, 675	**0.24**	**0.35**
45, 900	**0.29**	-0.06
45, 1200	**0.31**	*-0.14*
45, 1600	**0.25**	*-0.18*
45, 2150	**0.27**	*-0.23*
56, 500	**0.19**	**0.25**

56, 675	**0.19**	**0.26**
56, 900	**0.28**	0.09
56, 1200	**0.24**	*-0.14*
56, 1600	**0.25**	*-0.12*
56, 2150	**0.19**	*-0.23*
67, 500	**0.10**	**0.49**
67, 675	**0.20**	**0.27**
67, 900	**0.26**	0.02
67, 1200	**0.23**	-0.09
67, 1600	**0.24**	*-0.10*
67, 2150	**0.25**	*-0.23*
Experiment 4b		
Eigenvalues: **PC_1 (6.40)**, PC_2 (0.52), PC_3 (0.21)		
Variance explained	***79.59%***	***6.50%***
45, 500	*-0.17*	**0.43**
45, 675	*-0.25*	**0.26**
45, 900	*-0.25*	**0.15**
45, 1200	*-0.28*	-0.03
45, 1600	*-0.27*	*-0.18*
45, 2150	*-0.24*	*-0.28*
56, 500	*-0.20*	**0.22**
56, 675	*-0.24*	**0.14**
56, 900	*-0.26*	0.08
56, 1200	*-0.25*	*-0.16*
56, 1600	*-0.25*	*-0.20*
56, 2150	*-0.25*	*-0.24*
67, 500	*-0.09*	**0.33**
67, 675	*-0.18*	**0.22**
67, 900	*-0.24*	**0.26**
67, 1200	*-0.22*	0.09
67, 1600	*-0.25*	*-0.14*
67, 2150	*-0.29*	*-0.41*
Experiment 4c		
Eigenvalues: **PC_1 (7.01)**, PC_2 (0.70), PC_3 (0.32)		
Variance explained	***75.91%***	***7.53%***
45, 500	*-0.23*	*-0.36*
45, 675	*-0.27*	*-0.22*
45, 900	*-0.27*	0.05
45, 1200	*-0.27*	**0.13**
45, 1600	*-0.26*	**0.24**
45, 2150	*-0.22*	**0.25**
56, 500	*-0.20*	*-0.38*
56, 675	*-0.20*	*-0.36*
56, 900	*-0.24*	*-0.10*
56, 1200	*-0.25*	**0.14**
56, 1600	*-0.23*	**0.24**
56, 2150	*-0.22*	**0.23**
67, 500	*-0.15*	*-0.30*
67, 675	*-0.20*	*-0.17*
67, 900	*-0.24*	*-0.18*
67, 1200	*-0.26*	0.02
67, 1600	*-0.26*	**0.21**
67, 2150	*-0.25*	**0.28**
Experiment 4c (retest)		
Eigenvalues: **PC_1 (6.13)**, PC_2 (0.55), PC_3 (0.37)		
Variance explained	***75.96%***	***6.79%***
45, 500	*-0.24*	*-0.27*
45, 675	*-0.26*	*-0.16*

45, 900	*-0.27*	**0.15**
45, 1200	*-0.26*	**0.11**
45, 1600	*-0.25*	**0.27**
45, 2150	*-0.21*	**0.29**
56, 500	*-0.19*	*-0.40*
56, 675	*-0.19*	*-0.29*
56, 900	*-0.26*	*-0.23*
56, 1200	*-0.25*	**0.14**
56, 1600	*-0.26*	**0.21**
56, 2150	*-0.20*	**0.25**
67, 500	*-0.18*	*-0.36*
67, 675	*-0.22*	*-0.25*
67, 900	*-0.25*	*-0.16*
67, 1200	*-0.24*	**0.11**
67, 1600	*-0.26*	**0.15**
67, 2150	*-0.24*	**0.20**

Note. Results from Principal Component Analysis (PCA) for each experiment separately showing the first two PCs, the variance they explain, the eigenvalues for the first three PCs (those greater than 1 are in bold), and the loadings of the cell-level SI are shown. The signs on the loadings are arbitrary but their pattern of contrast is not; we use italic to highlight negative loadings and bold to highlight positive loadings, with an arbitrary threshold of ±0.10 for difference from 0 (regular font). Note that in these analyses we also include the retest data from Experiments 4a and 4c, treating them as separate experiments referred to here as "4a (retest)" and "4c (retest)".

Finding clusters of similar participants

In summary, we were able to provide a good characterization of participants' behaviour on the missing fundamental task using three summaries: the overall Schneider Index (SI), the Consistency Index (CI), and the order effect (OE). Each participant's behaviour can therefore be conceptualized as a point in a 3-dimensional space defined by SI×CI×OE and bounded between –1 and +1 on the first dimension, –1 and +1 on the second, and 0 and 1 on the third. A visual representation of all our participants is given in Figures 6A and 6B in the main paper, showing the two projections on the SI×CI and SI×OE. These same representations are shown here in color as Figures S1A and S1B.

To ascertain the apparent existence of groups of participants with similar behaviour, we conducted a *k-means clustering* analysis (Hartigan & Wong 1979). For a given number of groups, *k*, this method tries to allocate each participant to the group with the closest mean, based on the inter-participant distances. We computed these distances as the Euclidean distances between all pairs of participants in the 3-dimensional space SI×CI×OE; therefore participants with very similar behaviour will have very small distances between them, while participants who differ will be further away. However, we do not know *a priori* the optimum number of such clusters, *k*, and therefore we used an automatic search procedure which selects the best *k* on the basis of the *Calinski Harabasz index* (Calinski & Harabasz 1974; function `kmeansruns` in `R`'s library `fpc`), which optimizes the within- versus between-cluster distances.

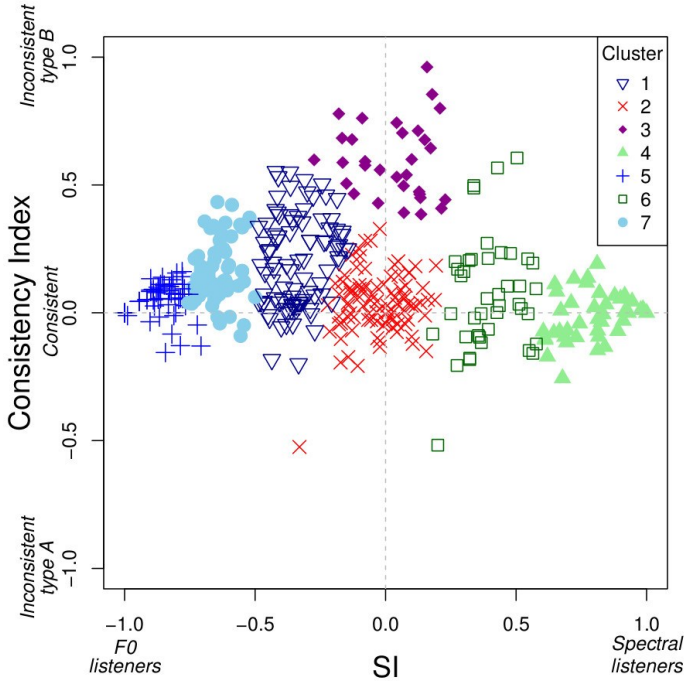

Figure S1A (Color version of Figure 6A in the main paper): Distribution of participants across all experiments in the SI×CI space, also showing the seven optimal clusters using symbols and colours. 'Inconsistent type A' participants give predominantly spectral responses to lower frequency stimuli and F0 responses to higher frequency stimuli; the 'Inconsistent type B' pattern of responses would be the reverse, but as can be seen this pattern is not found.

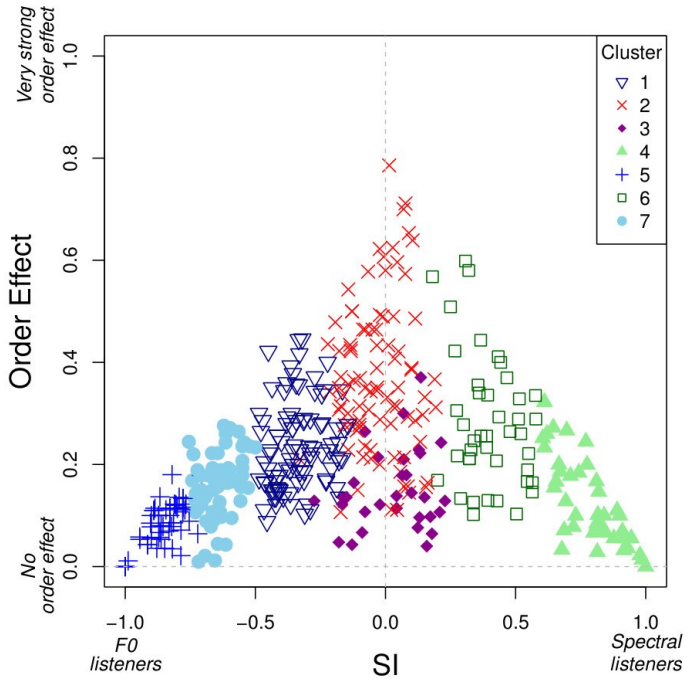

Figure S1B (Color version of Figure 6B in the main paper): Distribution of participants across all experiments in the SI×OE space, also showing the seven optimal clusters using symbols and colors.

We found that the optimal value of k is 7, and the clusters, as discussed in the main paper, are quite interpretable. To test the robustness of k=7, we noted that the Euclidean distance is a particular case (with p=2) of the general Minkowski distances, which for a pair of n-dimensional points $(x_i)_{i=1,n}$ and $(y_i)_{i=1,n}$ is defined as $(\sum_{i=1,n} |x_i - y_i|^p)^{1/p}$, where the order p is fixed. Thus, we repeatedly computed the optimal number of clusters k for different orders p, and we found that k=7 is robust for $2 \leq p \leq 6$, while for Manhattan distances (p=1), k=3, and for Euclidean distances (p=2) k=10 is equally good and suggest very similar clusters. The k=7 clusters obtained using the Euclidean distances are shown in Figure S1 (panels A and B).

Combining different experiments

As described in the main paper, we amalgamated the results from several experiments conducted at different times and in different places, with different participant groups and slightly different stimulus characteristics. This necessarily raises the question of the legitimacy of this procedure. Here we expand on the justifications given in the main paper.

As described in detail above, we conducted separate Principal Component Analyses for each experiment before amalgamating them. For all experiments we found similar patterns reflected by the first two principal components, not only in terms of structure but also in the amount of variance explained. This provides *a priori* support for the idea that the experiments reveal fundamentally similar patterns of behaviour. Impressionistic comparison of the distributions of the three data summaries derived above (SI, CI and OE) for the separate experiments also seems to confirm that the results of the individual experiments are quite similar, with the possible exception of 3a and 3b. These comparisons are illustrated in Figure S2 (panels A, B and C), in which the distributions are smoothed using Kernel Density Estimation (cf. Figure 5 in the main paper).

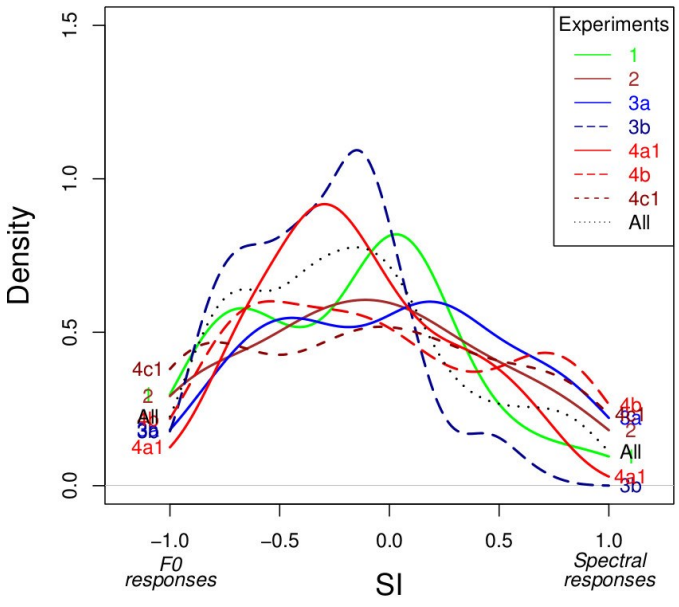

Figure S2A: Smoothed distribution of SI (overall Schneider Index) across experiments.

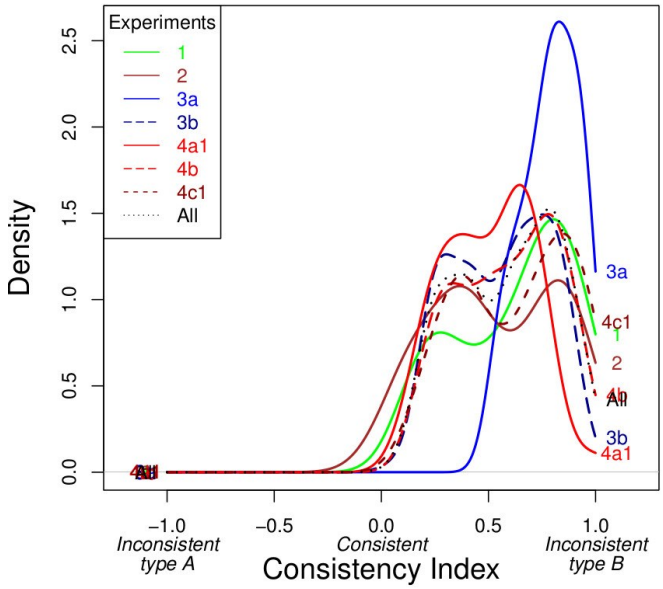

Figure S2B: Smoothed distribution of CI (Consistency Index) across experiments.

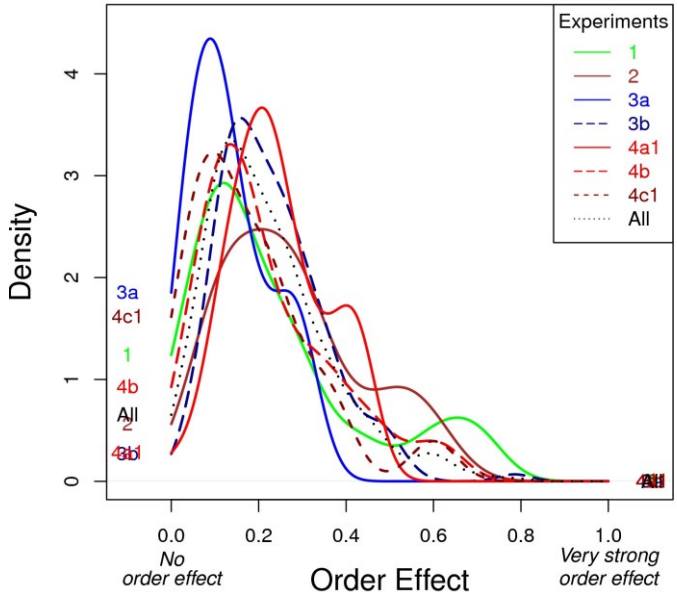

Figure S2C: Smoothed distribution of OE (Order Effect) across experiments.

To test more rigorously for differences between experiments, we conducted ANOVAs followed by pair-wise t-tests (corrected for multiple testing using Tukey's Honest Significant Difference (as implemented in **R** by the `aov` and `TukeyHSD` functions) as well as Kolmogorov-Smirnov tests (implemented in **R** by the `ks.test` function) corrected for multiple testing using Holm's (1979) method. The results of

these analyses are summarized in Table S2. Overall, there are few significant differences, confirming that the patterns across the experiments are very similar, and supporting the idea that the task is robust to small differences of methodology. At the same time, these tests confirm that 3a and to a lesser extent 3b are somewhat different, especially with respect to the Consistency Index and the Order Effect. We therefore conducted supplementary clustering analyses, first excluding Experiment 3a and then excluding both Experiments 3a and 3b. These analyses yielded a similar cluster structure, though the actual number of clusters varied: excluding 3a yielded 10 as the optimal number of clusters, while excluding 3b gave an optimum of 4. In both cases, however, 7 was also close to this optimum. These results suggest that the inclusion of Experiments 3a and 3b does not distort the conclusions reported in the main paper based on the amalgamated results.

Table S2

Comparison of individual experiments

Measure	ANOVA		Pair-wise *t*-tests			Kolmogorov-Smirnov	
	***F*(6,405)**	***p***	**Groups**	**Difference**	***p***	**Groups**	***p***
SI	3.96	0.0007	3b-4b	0.27	0.0014	3b-4b	0.0095
						3a-3b	0.017
						3b-4c1	0.0068
CI	4.96	6.53e-05	2-3a	-0.25	0.0088		
			3a-3b	0.23	0.00025	3a-3b	0.008
			3a-4a1	0.29	0.000028	3a-4a1	0.00059
			3a-4b	0.20	0.0075	3a-4b	0.048
OE	3.09	0.0057	2-3a	0.14	0.02		
			3a-3b	-0.09	0.045	3a-3b	0.031
			3a-4a1	-0.11	0.038	3a-4a1	0.039
						3b-4c1	0.041

Note. The pair-wise t-tests and Kolmogorov-Smirnoff tests are corrected for multiple testing using Tukey's HSD and Holm's procedures respectively.

References

Calinski, R. B., & Harabasz, J. (1974). A dendrite method for cluster analysis, *Communications in Statistics*, 3, 1-27.

Hartigan, J. A., & Wong, M. A. (1979). A *k*-means clustering algorithm. *Applied Statistics,* 28, 100–108.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* **6**:65–70.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Verlag: New York.