Psychomusicology: Music, Mind, and Brain
2015, Vol. 25, No. 2, 103–115

0275-3987/15/$12.00 http://dx.doi.org/10.1037/pmu0000085

Minding the Gap: An Experimental Assessment of Musical Segmentation Models

Joanna L. Hutchison
University of Texas at Dallas

Timothy L. Hubbard
Texas Christian University

Nicholas A. Hubbard, Ryan Brigante, and Bart Rypma
University of Texas at Dallas

Three experiments examined the perceptual validity and parsimony of 2 musical segmentation models—the perceived phrase structure (PPS) model (Temperley, 2001) and the generative structural grammar of music (GSGM) model (Lerdahl & Jackendoff, 1983). Each of these models is composed of a set of rules that specifies where the perception of a natural break in musical phraseology is most likely to occur. In 3 experiments assessing the models from different perspectives, participants listened to melodies and then drew a vertical line in a row of equally spaced dots corresponding to where a natural break in the stimulus was heard, with each dot representing 1 note regardless of the duration of the note (cf. Deliège, 1987). The PPS and GSGM models were similar in their overall predictive accuracy; however, the PPS model gap rule, in which large temporal intervals between notes yield the perception of a musical phrase boundary, explained nuances arising within the broader GSGM model. The PPS gap rule is therefore suggested as a strong factor in musical segmentation, adding to our understanding of the perception of musical phrases.

Keywords: segmentation, parsing, music

Music and language are parallel in many respects (for overviews, see Besson & Friederici, 1998; Patel, 2010), including their origins and historical development (Wallin, Merker, & Brown, 2000), patterns of development across the life span (Drake, 1998), and elements of cognitive processing (Swinney & Love, 1998). Although music and language are different types of stimuli and utilize different cognitive operations (Patel, 1998; Peretz, 1993; Peretz et al., 1994), they appear to share several structural properties (Blacking, 1973; Krumhansl, 1998; Lerdahl & Jackendoff, 1983; Raffman, 1993; Sloboda, 1985). For example, in music and in language, structural rules specify how combinations of smaller units form larger structural frameworks (e.g., notes forming melodies, words forming sentences). However, structural properties of music are more ambiguous and fluid than are structural properties of language (Jackendoff & Lerdahl, 1981; Jescheniak, Hahne, & Friederici, 1998), and the specific structural rules by which a listener might segment a larger musical stimulus into smaller meaningful units are not entirely clear. To better understand the process of musical segmentation, the experiments presented here provide an empirical evaluation of structural rules suggested by two different models of music cognition: the perceived phrase structure model (PPS; Temperley, 2001) and the generative structural grammar of music model (GSGM; Lerdahl & Jackendoff, 1983).

Based in part on the notion of generative structural grammars in language (Chomsky, 1965, 1968, 1957/2002; Yngve, 1960), analogous generative structural grammars have been proposed for music (Deutsch & Feroe, 1981; Lerdahl & Jackendoff, 1983; Tenney & Polansky, 1980; Winograd, 1968). The unit of the motif, which can be as small as two notes, is important both structurally and psychologically (see Jackendoff & Lerdahl, 1981; Zbikowski, 1999). Generative grammars combine these motifs into basic phrase units to form the lower level of a hierarchical structure (see Gregory, 1978; Johnson, 1965; Rosenberg, 1968, for empirical support for the functional use of such a structural grammar). Given the fundamental importance of the phrase, and the importance of understanding basic units before proceeding to more complicated hierarchies based on combinations of those basic units, the focus here will be upon segmentation within a local context (i.e., on phrase-level context rather than on extended and hierarchically organized context). To wit, we assessed how participants determine phrase boundaries in musical stimuli and segment those musical stimuli into meaningful parts. There is a paucity of literature assessing segmentation models in diverse (i.e., experienced and inexperienced listener) human (i.e., nonmachine) populations. The relative novelty of our study lies within the populations that

Joanna L. Hutchison, Department of Behavioral and Brain Sciences, University of Texas at Dallas; Timothy L. Hubbard, Department of Psychology, Texas Christian University; Nicholas A. Hubbard, Ryan Brigante, and Bart Rypma, Department of Behavioral and Brain Sciences, University of Texas at Dallas.

We thank Nancy Rowe at the University of Texas at Arlington for her assistance with the statistical analyses. We would also like to thank Jon Courtney for his assistance in conducting some of the experimental sessions.

Correspondence concerning this article should be addressed to Joanna L. Hutchison, Department of Behavioral and Brain Sciences, University of Texas at Dallas, Center for Brain Health, 2200 W. Mockingbird Ln., Dallas, TX 75235. E-mail: joanna.hutchison@utdallas.edu

we studied and the hypotheses that we tested. Specifically, we studied humans with a range of musical experience and compared their segmentation responses with theoretical segmentation points generated by two prominent musical segmentation theories.

The GSGM model is a prominent and influential theory that proposes several grouping preference rules for determining the locations of local context phrase boundaries in music (see Table 1). The GSGM model suggests that when one or more of these grouping rules is fulfilled, then perceived segmentation (i.e., perception of a phrase boundary) typically occurs. The rules are not mutually exclusive, and different rules can be combined to create different strengths or levels of segmentation. Deliège (1987) examined the extent to which perceptual grouping corresponded with the local context segmentation rules of GSGM for a group of listeners. Deliège presented short musical excerpts, and musician or nonmusician participants indicated the point at which they heard natural breaks in the music by drawing a vertical line in a row of horizontally oriented dots in which successive dots represented successive notes. Results were interpreted as supporting the validity of GSGM rules for both participant groups. Deliège further noted that segmentation often occurred one note later than predicted in the case of the articulation and length change rules and that "sensitivity to a sensation of a gap in music perception may be considered, by the way, as a key element in the grouping behavior" (p. 343).

Peretz (1989) examined potential segmentation rules of pitch skip (similar to the GSGM register change rule), temporal pause (similar to the GSGM length rule), and pause plus parallelism (i.e., sounds on either side of a phrase boundary were of the same duration, and the two musical phrase groupings were of the same duration such that durational symmetry existed). Using the same methodology as Deliège (1987), Peretz found that perceptual boundaries indicated by musicians and by nonmusicians were in high agreement with the boundaries predicted by the proposed rules. In a similar study, Frankland and Cohen (2004) asked musically trained or untrained participants to press a button during presentation of musical pieces to indicate perceived natural breaks in the music, and they found that GSGM attack and length rules operated similarly and could potentially be combined on the basis of parsimony. Frankland and Cohen's conclusion parallels the conclusion of Deliège (1987) that an overarching type of length rule (involving the number of notes and the duration of the phrase) encompasses several GSGM rules for both musicians and nonmusicians.

Temperley (2001) combined some of the GSGM local context segmentation rules and hierarchical segmentation rules, and further suggested that an optimal phrase length contributes to the perception of musical segmentation. The result was a set of three rules suggested to determine perceived phrase structure (PPS; see Table 1). In the PPS model, the gap rule is postulated to be the primary segmentation rule, with the phrase length and parallelism rules modulating the segmentation suggested by the gap rule. Ultimately, the perception of segmentation is based upon the relative strengths of these three factors.

Pearce, Müllensiefen, and Wiggins (2010) reviewed several models of musical grouping and segmentation, and they also proposed a novel model based on statistical learning and information dynamics which they referred to as Information Dynamics of Music (IDyOM). Within the IDyOM model, the listener's learning-based expectations drive the perception of musical phrase boundaries. Pearce et al. compared performance of the following models and rules based on how well their predictions agreed with expert-generated phrase markings in folk tunes: IDyOM; PPS (referred to as the "Grouper" model); Local Boundary Detection Model (Cambouropoulos, 2001), which is a quantification of GSGM rule magnitude; four local-context GSGM rules, as quantified by Frankland and Cohen (2004); and two language-based, statistical learning models—the Pointwise Mutual Information model (Brent, 1999) and the Transition Probabilities model (Saffran, Aslin, & Newport, 1996). The GSGM rest rule was the most predictive of all of the models and rules, and of particular relevance to the current investigation. Further, Pearce et al. noted that all of the top-performing models, including the PPS model and

Table 1
Summary of the Theoretical Structure of the GSGM and PPS Models

GSGM model (Lerdahl & Jackendoff, 1983)	PPS model (Temperley, 2001)
Grouping rule 1: Length requirement Groupings should contain more than one note	Segmentation rule 1: Gap Large inter-onset intervals (IOIs) and large offset-to-onset intervals (OOIs) result in perception of a phrase boundary
Grouping rule 2: Proximity Note proximity affects grouping a. Rest—Notes separated by a slur or a rest will be divided into different groupings b. Attack—A longer note surrounded by shorter notes gives a feeling of finality, leading to the perception of a phrase boundary after the longer note	Segmentation rule 2: Phrase length Listeners prefer phrases approximately eight notes in length
Grouping rule 3: Change Any type of change in the music yields segmentation at the point of change a. Register—A jump of an octave or greater between successive notes b. Dynamics—Changes in volume c. Articulation—Changes in the sustained articulation style of notes d. Length—Changes in note duration e. Timbre/Instrumentation—Changes in the instrument or sound in use	Segmentation rule 3: Parallelism Phrase boundaries are more likely to be perceived if the successive phrases begin at metrically parallel points in the musical structure (i.e., on the same beat of the measure)

their own IDyOM model, incorporated rests or temporal gaps in their estimation of boundary placement.

Bruderer, McKinney, and Kohlrausch (2012) suggested that the segmentation of musical boundaries could be predicted by a combination of timbre change, onset (cf. Cambouropoulos, 2001), and the rest rule (as developed by Frankland & Cohen, 2004; based on Lerdahl & Jackendoff, 1983). Bruderer et al.'s operationalizations of the Local Boundary Detection Model onset rule (Cambouropoulos, 2001; i.e., larger interval sizes are more likely to be perceived as segmentation boundaries) and the rest rule (i.e., longer rest durations are more likely to be perceived as segmentation boundaries) are similar to the notion of gap size in the PPS model. Although Bruderer et al. did not evaluate the PPS model, they suggested that the PPS model should be considered in future segmentation research. Accordingly, the experiments reported here examined the extent to which the PPS and GSGM models of musical segmentation predict perceived segmentation of melodies.

In Experiments 1–3, participants with a range of musical experience were presented with a set of musical stimuli and were asked to indicate the strongest segmentation point that they heard within each melody. PPS and GSGM models were used to predict a "theoretical" segmentation point within each melody. In Experiments 1 and 2, novel melodies were created to contain such a point for the PPS and GSGM models, respectively; these segmentations represented a "best-case" scenario for each model. By creating stimuli that incorporated the rules of the model in a clear manner, we expected to find that segmentation responses would correspond to a great degree with the factors of the models, if the models were accurate in their predictions. In Experiment 3, stimuli consisted of melodies drawn from the sight-singing literature that contained segmentation points suitable to both the PPS and GSGM models. The validity of the PPS and GSGM models was judged by the degree to which participants' responses matched the theoretical segmentation points predicted by each model. Data were analyzed using generalized linear models, treating the rules as factors in the analyses. We assessed whether including each factor (i.e., rule) improved the fit of the model in predicting whether participants chose the theoretical segmentation point.

Experiment 1: Testing the PPS Model

In Temperley's (2001) initial report, the PPS model was used to generate musical phrase segmentations on short folk songs and sight-singing materials that were comparable with musical segmentations made by a group of composers (i.e., very experienced listeners). However, segmentation predictions suggested by the PPS model have not yet been explicitly compared with segmentation responses of less musically experienced (i.e., average) listeners. If segmentations predicted by the PPS model also correspond to segmentations made by less musically experienced listeners, then the PPS model might be the most generalizable and parsimonious of the formalized accounts of musical segmentation available.

Experiment 1 compared predictions of the PPS model with perceived segmentations in a diverse group of listeners. Given that low-level construction of a musical percept is dissociable from musical experience (Peretz, 1989), and that patterns of responding are often similar for musicians and for nonmusicians (although nonmusicians' responses are often less consistent than those of musicians; Deliège, 1987; Frankland & Cohen, 2004; Peretz, 1989), we predicted a high level of correspondence (i.e., alignment) between segmentation responses of a musically unselected sample of human listeners and PPS model predictions. If this were to hold true, then the same rules would be used by those with less experience, although possibly to a lesser degree (i.e., more variability in segmentation, but adhering to the same average segmentation pattern), or only the most robust segmentation rules would be used (i.e., strong overall segmentation, but only reflecting a subset of the strongest segmentation rules).

Method

Participants. Seventeen individuals (6 male, 11 female) were selected to participate from the students, faculty, and other individuals associated with Texas Christian University; these individuals were primarily traditional collegiate undergraduates. Recruitment was not based on musical expertise, although we obtained information regarding the number of years of musical experience (playing an instrument or singing, participation in a band or chorus, enrollment in music lessons) from each participant (see section on *Preliminary analysis* below). Participants who were students received partial course credit for participation. Other participants received monetary compensation for their time.

Apparatus. Stimuli were presented using a Compaq Pentium IV Personal Computer with the Microsoft Windows XP operating system, at a sound level of ~57 dB (determined by a Radio Shack Digital Sound Level Meter, Cat. No. 33–2055). Koss UR-15C headphones were used for stimulus presentation; a built-in pad encircled the speakers such that outside noises were dampened and comfort was maximized. Cakewalk Sonar Studio 3.0 software was used to generate stimuli in MIDI format, which were then recorded onto a stereo audio track using the acoustic piano sound patch. Audio tracks were then exported in .mp3 format into Microsoft PowerPoint for presentation to participants.

Stimuli and response materials.

Auditory stimuli. Novel musical stimuli were designed to test the PPS model rules of gap size, phrase length, and parallel construction (see Figure 1 for examples). Stimuli were initially

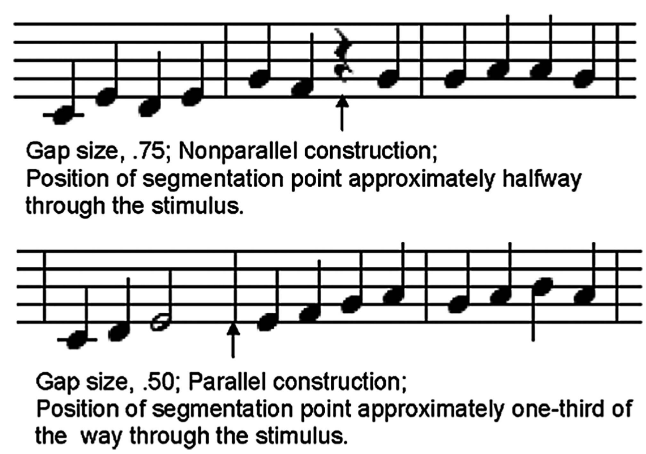

Figure 1. Examples of stimuli used in Experiment 1, created based on PPS model (Temperley, 2001) segmentation rules. Arrows indicate points of theoretical segmentation.

generated in the key of C major and were then duplicated in the key of F major, and this resulted in a total of 110 stimuli. By presenting stimuli in two different musical keys, the potential confound of pitch height and harmonic function across stimuli was diminished. The tempo was 100 beats per minute, and each beat was equivalent to one quarter note in the music (i.e., the quarter note received the beat). All notes were quarter notes except for those notes made longer or shorter by the experimental manipulation; all longer notes were half notes, and shorter notes were eighth, 16th, or 32nd notes. Two blocks of trials were presented. Three different random stimulus orders were used in each block, with the first block and second block counterbalanced for each stimulus order, leading to a total of six different possible stimulus orders.

Gap size. Following Temperley (2001), gap size was defined as the proportional length of a note plus two times the proportion of the measure taken up by any rest immediately following that note. This formula placed more emphasis on the rest portion of the IOI because of the assumption in the PPS model that rest is a stronger grouping cue than is proportional note length; see Temperley for further explanation of gap score derivation. For example, a quarter note (.25 measures in duration, with 1 beat equivalent to .25 of a 4-beat bar) followed by a half rest (.50 measures in duration, with 1 beat equivalent to .25 of a 4-beat bar) would have a gap score of 1.25—that is, $.25 + (.50 \times 2) = 1.25$. We were able to create varying gap scores across stimuli by manipulating various aspects of the stimuli: varying rest lengths (half, quarter, eighth, 16th, 32nd rests), changes in note length (e.g., a series of shorter notes followed by a series of longer notes), and changes in articulation (i.e., the sustained shortening or lengthening of notes); additional examples of gap score calculations under conditions of different stimulus manipulations are shown in Figure 6. Gap scores were generated for each stimulus, with the locally highest gap score within a stimulus constituting the predicted segmentation point of that stimulus.

Phrase length. As previously indicated, the PPS model hypothesizes an optimal phrase length of 8 notes. The modal stimulus length for our stimuli was 12 notes; however, some of the manipulations required either shorter or longer stimuli. This resulted in stimuli ranging in length from 6 to 17 notes. In order to provide variability in segmentation points across stimuli and to provide sufficient melodic material on both sides of each potential phrase boundary, we placed theoretical phrase boundaries ~1/3, 1/2, or 2/3 of the way through each stimulus. Segmentation was expected to occur most frequently at the 1/3 and 2/3 theoretical phrase boundaries because an eight-note phrase typically resulted from either of those segmentation points (either at the beginning or the end of the excerpt). Boundary placement was approximate such that segmentation was not expected at precisely the same points within each stimulus, and this facilitated model analysis by allowing for a categorical assessment of the accuracy of the contention that the eight note phrase is the most optimal length. If segmentation were to only occur upon encountering eight notes, then the 2/3 theoretical phrase boundary would be expected to result in the greatest amount of segmentation in contrast to the 1/3 phrase boundary.[1]

Parallel construction. Stimuli were constructed with parallel phrase structure (i.e., with theoretical phrase boundaries beginning on the same beat of the measure) or without parallel phrase structure (i.e., with theoretical phrase boundaries beginning on different beats of the measure; see Figure 1 for examples of parallel phrase construction and nonparallel phrase construction). Stimuli constructed with parallel phrase structure were expected to result in increased response/model correspondence (i.e., result in a larger alignment between segmentation responses of the participants and predictions of the model).

Theoretical segmentation points. The PPS model does not explicitly specify how the gap, phrase length, and parallelism rules interact; instead, Temperley (2001) generated a computer algorithm to obtain the most accurate responses (i.e., responses according to phrase markings by professional composers). However, it was inappropriate for us to use Temperley's (2001) computer model, which examines segmentation based on the model as a singular entity, because we were interested in testing all three of the PPS model rules both separately and in concert. Given the observation that GSGM rest and attack rules predominate unless those rules are ambiguous or absent (Lerdahl & Jackendoff, 1983), the following operationally defined weightings were used: If gap size (see *Gap size*, above) was medium (i.e., .50 to .99) or large (i.e., 1 or greater), theoretical segmentation points were based solely on gap score, whereas if gap size was small (i.e., less than .50), theoretical segmentation points were determined based on the convergence of two or more rules. For example, if a gap was small and phrase length and parallel construction suggested phrase boundaries at two different points, then the theoretical segmentation point was based solely upon gap. However, if phrase length and parallel construction converged to suggest a single segmentation point, then the small gap was ignored in favor of the segmentation point suggested by the other two rules. Thus, segmentation was operationalized in terms of both the magnitude and convergence of the PPS rules.

Response materials. Following Deliège (1987), each auditory stimulus was represented on a paper response sheet by a separate horizontal row of equally spaced and equally sized dots, with one dot representing each note; differences in note duration or temporal gaps were not represented on the response sheet. The Musical Background Questionnaire consisted of questions regarding musical experience, including the type, level, quality, and duration of musical experiences and training.

Procedure. After explanation of the procedure and obtaining signed informed consent, participants were presented with three examples of auditory stimuli similar to those used in the experimental trials, followed by three practice trials, and were then given

[1] Meter and the total passage of time would seem to be important in determining how many notes should constitute a phrase—an issue that was partially addressed by the PPS model (Temperley, 2001) with regard to the PPS gap and parallelism rules. The PPS model accounts for the total passage of time in the gap rule by taking into account the proportion of the measure taken up by a note (adjusted for rests), but no such accommodation is made for the phrase length rule. The PPS model acknowledges parallelism and meter to be dependent on one another, but the model does not directly address this dependency issue with regard to the phrase length rule. Testing the PPS model, we did not incorporate meter or total passage of time as additional factors into the model; however, all of the stimuli in Experiment 1 were created in the same meter, and were recorded at the same speed. As previously mentioned, quarter notes predominated the stimuli, although notes of varying lengths were required to better examine gap size.

an opportunity to ask procedural questions. Two blocks of trials followed. Each block contained a pseudorandomized, equal number of trials representing each trial type (i.e., convergence of model rules in various combinations); the break between blocks was offered to give participants a chance to rest. On each trial, participants listened to three presentations of the auditory stimulus, and on the third presentation, participants indicated where they heard the strongest natural break in the music by drawing a single vertical line on the response sheet within the horizontal row of dots for that trial. Participants pressed the spacebar to initiate presentation of each trial, and they were allowed to progress at their own pace (with one exception: a mandatory break of at least 3 min between the end of the first block and the beginning of the second block). After completing the segmentation task, participants filled out the Musical Background Questionnaire and were then debriefed and thanked.

Results and Discussion

For each trial, participants' responses were scored as 1 or 0, reflecting a match or nonmatch, respectively, with the theoretical segmentation point derived from the PPS model. Data were analyzed using generalized linear mixed models (GLMMs) incorporating random effects using the PROC NLMIXED (i.e., nonlinear, mixed design) procedure in the Statistical Analysis Software (SAS) programming package (see Agresti, 2002, for an overview of the statistical treatment of categorical data, as well as an overview of specific SAS procedures; see also Amemiya, 1981; SAS Institute, 2003). Within the nonlinear mixed effects model, each level of each factor was specified separately and beta weight estimates were obtained for each level of each factor; a singular beta weight for each factor was not appropriate given that the effects were not necessarily linear.

Preliminary analysis. A common practice in the music cognition literature has been to classify experimental participants as "trained/experienced" or as "untrained/inexperienced," and a cutoff point of three years of training/experience has often been used (Bartlett & Dowling, 1980; Dowling, 1984; Verheul & Geuze, 2004). However, in Experiment 1 participants' self-reported years of musical experience were used to classify each participant into one of three groups, with low musical experience defined as less than 3 years of musical experience ($n = 6$), intermediate experience defined as 3 to 9 years of experience ($n = 9$), and high musical experience defined as 10 or more years of musical experience ($n = 2$).[2] We chose to utilize the three-way classification rather than the more traditional two-way classification for two main reasons. First, many participants in the intermediate experience group were not clearly either musically trained or not musically trained (e.g., many of these participants indicated that they had played an instrument for three or four years in a school band or orchestra but no longer played that instrument, and some of these participants commented during debriefing that they did not consider themselves to be musically trained). Second, the three-way classification could potentially afford a clearer picture of the data, and follow-up studies could be conducted with recruitment based on musical experience if warranted by the patterns in the data. However, because of the small sample size in the high experience group, the data were also analyzed with the more traditional two-way experience classification, with the cutoff as three years of experience (inexperienced, $n = 6$; experienced, $n = 11$). In both analyses, there was no effect of musical experience on participants responding, $ps > .05$.

There were no effects of stimulus presentation key on segmentation responses, $ps > .05$. An unexpected practice effect was found in the data, with higher response/model correspondence during the second block of trials (90.70%) than during the first block of trials (86.63%; $\beta = .41$, $p < .05$). It might be that a longer practice trial period (our participants received three practice trials) is necessary for participants to acclimate properly to the experimental procedure and perform at optimal levels (but see Experiment 3).

PPS rules. As shown in Figure 2, there was high response/model correspondence (i.e., agreement of participants' responses with model predictions), as participants' segmentation choices corresponded with predictions based on the PPS rules for 88.66% of responses. In contrast, chance-level responding, which was a function of the number of notes in a stimulus, varied from 5.9% to 16.7%. Therefore, segmentation responses corresponded with PPS rules at above-chance levels.

Gap size. Gap size was examined as a categorical variable (i.e., small, medium, or large, with gap sizes of $< .5$, .5 to .9, and ≥ 1, respectively; Figure 2). Both small gaps ($\beta = -2.67$, $p = .0004$) and medium gaps ($\beta = -1.95$, $p = .0047$) led to less response/model correspondence compared with large gaps. Count R-squared (Long & Freese, 2006; also see Amemiya, 1981, and Windmeijer, 1995), a measure of goodness-of-fit indicating the proportion of responses predicted correctly by the model, was .89. Data were further examined using raw gap scores as a continuous variable to determine whether overall gap magnitude was predictive of segmentation. This was indeed the case, with larger gap magnitudes being more likely to result in segmentation responses, $\beta = 3.18$, $p < .0001$.

Phrase length. Theoretical phrase length was likewise examined as a categorical variable (i.e., the break was ~1/3, 1/2, or 2/3 of the way through the stimulus). The percentage of participant responses that agreed with the theoretical segmentation points was similar for all phrase lengths (see Figure 2), and phrase length did not interact with other variables, $ps > .05$.

Parallel construction. The parallel designation was categorical (i.e., parallel, not parallel). Parallel construction did not influence segmentation responses (Figure 2; but see the section *GSGM rules as substitutes for the gap rule* below), nor did parallel construction interact with the other variables, $ps > .05$.

Interim summary and rationale for further analyses. The only significant predictor of segmentation responses in the PPS model was gap size. This factor accounted for 88.66% of segmentation responses. Given the similarity of the PPS gap rule to the GSGM Proximity rest and attack rules, and Change articulation and length rules (cf. Table 1), an additional analysis was conducted to examine whether using these GSGM model rules could substitute for the gap rule. This analysis offered an evaluation of possible differences between the GSGM rules and the gap rule, and further assessed the overall validity of these variables in predicting musical segmentation responses.

[2] Assumptions of mixed effects models accommodate unequal group sizes, and therefore this is not a concern.

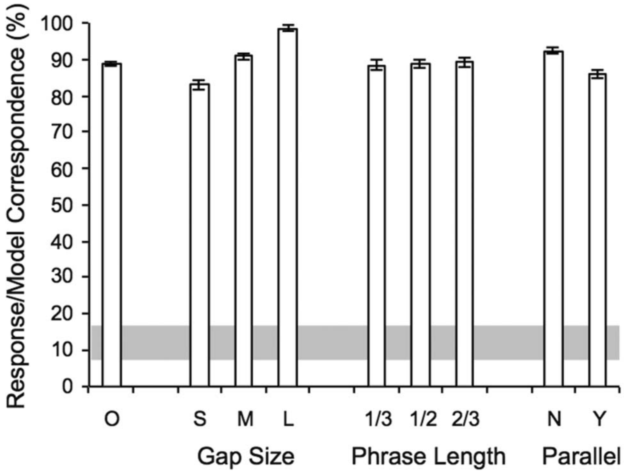

Figure 2. The correspondence of the PPS model (Temperley, 2001) with participant responses in Experiment 1. "O" refers to an overall score collapsed across gap size, phrase length, and parallel structure. "S," "M," and "L" refer to small, medium, and large gap sizes. "1/3," "1/2," and "2/3" refer to whether segmentation was predicted 1/3, 1/2, or 2/3 through the stimulus. "N" and "Y" refer to whether parallelism was not present (i.e., "no") or was present (i.e., "yes"). The horizontal gray bar reflects the range of chance performance. Error bars reflect standard error of the mean.

GSGM rules as substitutes for the gap rule. As seen in Table 2, response/model correspondence was high with GSGM rest and attack rules. Half, quarter, eighth, 16th, and 32nd rests all led to greater levels of segmentation behavior over baseline (with a base model that assumed length changes—i.e., the GSGM length rule). The gap rule can be construed as accounting for the rest rule; even 32nd (i.e., very small) rests were heard as natural breaks. Additionally, response/model correspondence was high with the GSGM attack rule. Given that the attack rule results in an altered gap size between notes (Deliège, 1987), the attack rule could be suggested to be a special case of a more general gap rule. Thus, the PPS gap rule more parsimoniously accounted for participants' patterns of responses than did the combination of rules from the GSGM model. Indeed, our analyses showed equivalent count R-squared values for the two models (one utilizing GSGM rules and the other using GSGM rules with the PPS gap rule substitution), suggesting that these models explain gap perception in music.

Counter-theoretic implications for models. When using GSGM rules as a substitute for the PPS model gap rule within the PPS model, we encountered two counter-theoretic results. First, the negative relationship between the articulation rule and response/model correspondence (see Table 2) was unexpected. One possibility is that articulation changes were smaller in scope than were length changes (i.e., articulation changes led to smaller gap sizes), and thus articulation changes led to a less salient or less natural break in the music. Consistent with the PPS gap rule, gap size mattered, with larger gaps more likely to produce higher response/model correspondence. Second, nonparallel phrase construction yielded greater response/model correspondence than did parallel phrase construction. Because of the complete crossing of the three PPS model rules of gap size, phrase length, and parallel construction in the stimuli, a short note following a rest might be perceived as a "pick-up" note and thus might have been construed as beginning a new phrase. Thus, the negative relationship between parallel construction and segmentation responses might result from experimental artifact. This possibility will be addressed with the use of folk tunes and other music drawn from the sight-singing literature in Experiment 3.

Experiment 2: Testing the GSGM Model

Data from Experiment 1 suggested that the gap rule from the PPS model might encompass multiple rules of the GSGM model, and so the PPS model might be more parsimonious than the GSGM model. However, the stimuli in Experiment 1 were specifically tailored to fit the PPS model, and so the results of Experiment 1 might reflect the use of artificial stimuli rather than a broader parsimony of the PPS model. To ensure that the results of Experiment 1 were not due to idiosyncrasies of stimuli specifically constructed to test the PPS model, Experiment 2 used the same methodology as Experiment 1, but the musical stimuli were specifically tailored to fit the GSGM model. If the PPS model is more parsimonious than the GSGM model, then response/model correspondence would either be lower with the GSGM model, or equivalent but require the use of more factors in specifying the model.

Method

Participants. Seventeen participants were recruited from the same participant pool used in Experiment 1, and none had participated in that experiment. Data from one participant was deleted for failure to follow instructions, leaving a total of 16 participants (4 male, 12 female). (For information regarding musical experience of the participants, see section on *Preliminary analysis*, below.)

Apparatus. The computing/headphone apparatus was the same as in Experiment 1.

Stimuli and response materials.

Auditory stimuli. Novel musical stimuli were designed to directly investigate the rest, attack, register, length, articulation, timbre, and dynamics rules of the GSGM model. The slur portion

Table 2
Data From Experiment 1 With Gap Re-Analyzed in Terms of Lerdahl and Jackendoff's (1983) Local Context Rules, Maintaining the Remainder of Temperley's (2001) Model

Rule	β	*p*
Lerdahl and Jackendoff's (1983) rules:		
Rest: Half rest	3.29	<.0001
Rest: Quarter rest	2.48	<.0001
Rest: Eighth rest	2.24	<.0001
Rest: 16th rest	1.87	<.0001
Rest: 32nd rest	1.36	.0002
Attack	2.48	<.0001
Articulation	−0.52	.0278
Temperley's (2001) rules other than gap:		
Nonparallel	0.81	.0003
Phrase length	*ns*	*ns*

Note. *ns* = not significant.

of the GSGM rest/slur rule was not addressed, as determination of when a slur occurred did not afford clear-cut experimental manipulation (see Frankland & Cohen, 2004). Example stimuli are provided in Figure 3. The default stimulus length was 12 notes; however, a shorter length was required in some conditions, resulting in stimuli ranging in length from 8 to 12 notes. As in Experiment 1, all stimuli were presented in the key of C major and in the key of F major. Two different timbres, acoustic piano and French horn, were used, and each stimulus was presented in each timbre (for stimuli testing the timbre rule, this was in the form of which timbre was presented first, as both timbres were available for each musical excerpt). All notes were quarter notes other than notes that were longer as required by the experimental manipulation (the longer notes were all half notes). There were a total of 108 stimuli.

Response materials. The response materials were the same as in Experiment 1, including the paper response sheet and the Musical Background Questionnaire.

Procedure. The procedure was the same as in Experiment 1.

Results and Discussion

Participants' responses were scored as 1 or 0, reflecting a match or nonmatch, respectively, with the theoretical segmentation point derived from the GSGM model. As in Experiment 1, data were analyzed using GLMMs incorporating random effects using the PROC NLMIXED (i.e., nonlinear, mixed design) procedure in the SAS programming language.

Preliminary analysis. Using the same three-way musical experience classification as in Experiment 1, each of the 16 participants was initially categorized as low experience ($n = 12$), intermediate experience ($n = 3$), or high experience ($n = 1$). Given that only one participant was categorized as high experience in the three-way classification, analysis of musical experience was limited to use of a two-way musical experience classification (inexperienced, $n = 12$; experienced, $n = 4$). As in Experiment 1, musical experience was not a significant factor in predicting segmentation responses; similarly, key and practice were not significant determinants of segmentation responses, $ps > .05$.

Although PPS model phrase length was not technically part of the GSGM rules under evaluation, we included it in the model (as 1/3, 1/2, and 2/3 theoretical segmentation points) to account for possible midstimulus response bias (cf. Deliège, 1987). Response alignment with 1/2 theoretical segmentation points was less than response alignment with 1/3 and 2/3 theoretical segmentation points ($\beta = -.47$, $p = .0089$), obviating this concern and lending support for the PPS phrase length rule. Consistent with the PPS model, we found that 1/3 and 2/3 boundaries resulted in higher response/model correspondence than 1/2 boundaries. This model yielded a Count R-squared of .81.

GSGM rules. As shown in Figure 4, 69.49% of the responses were consistent with predictions of the GSGM model (chance responding was 8.3% to 12.5%). Rests ($\beta = 1.65$, $p < .0001$), attacks ($\beta = 1.32$, $p = .0003$), and register jumps ($\beta = .61$, $p = .0254$) led to significant increases in response/model correspondence, whereas articulation ($\beta = -1.30$, $p < .0001$) and length ($\beta = -1.60$, $p < .0001$) led to significant decreases in response/model correspondence. Count R-squared was .79.

Delayed segmentation. Deliège (1987) and Frankland and Cohen (2004) observed that one-note delayed segmentation sometimes occurred with regard to GSGM rules. To examine whether a similar delayed segmentation occurred in Experiment 2, the data were reanalyzed to include a one-note delay for the articulation and length rule stimuli only (henceforth *plus-one scoring*). With this alteration, response/model correspondence was similar to that without a one-note delay (69.49% and 68.15%, respectively). However, if both the exact theoretical segmentation point and the theoretical segmentation point at a one-note delay were scored as consistent with the model (henceforth *expanded scoring*), response/model correspondence increased to 81.03%. This value constitutes a greater increase in model-consistent responding than occurred by chance when doubling the number of answers counted as correct (chance responding of 16.6% to 25.0%).

To further examine the delayed segmentation in Experiment 2, the length and articulation stimuli were coded for long-short transitions (i.e., longer-duration notes followed by shorter-duration notes) or short-long transitions (i.e., shorter-duration notes followed by longer-duration notes). Segmentation responses were assessed using exact scoring, plus-one scoring, and expanded scoring. As shown in Figure 5, for length rule stimuli, the plus-one scoring was not very beneficial with long-short transitions (7.81%), but was more beneficial with short-long transitions (59.38%). For articulation rule stimuli, the opposite was true: long-short transitions received a larger benefit (47.88%) than did short-long transitions (20.31%). Overall, this resulted in increased response/model correspondence if expanded scoring was utilized, with articulation response/model correspondence of 82.82% (short-long transitions) and 88.28% (long-short transitions), and length response/mode correspondence of 83.59% (short-long transitions) and 85.94% (long-short transitions).

Application of the gap rule. Frankland and Cohen (2004) suggested that the GSGM length and attack rules could be combined, and they stated "if the interval of time between the attack

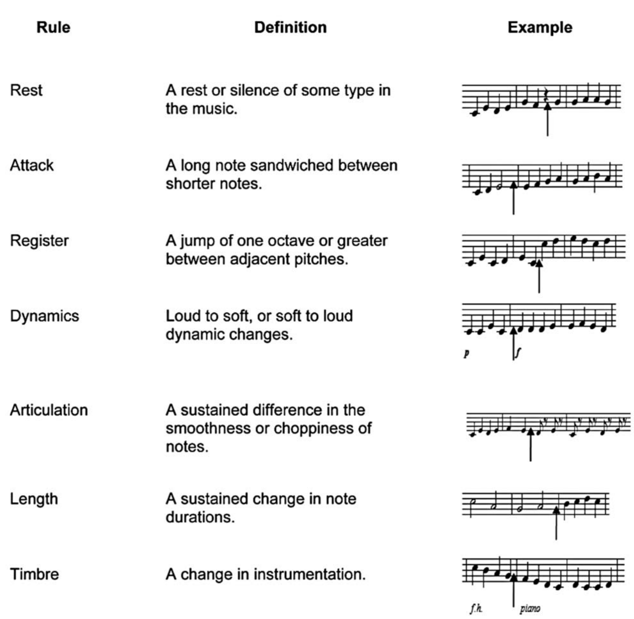

Figure 3. Examples of stimuli used in Experiment 2, created based on the GSGM model (Lerdahl & Jackendoff, 1983) local context segmentation (i.e., grouping) rules. Arrows indicate points of theoretical segmentation.

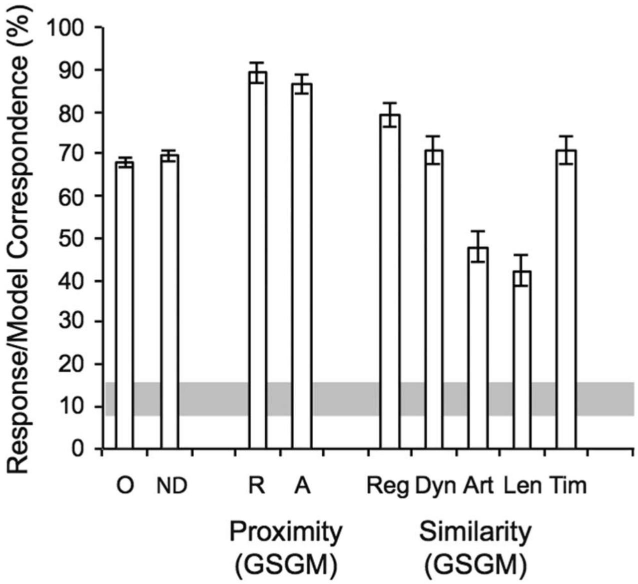

Figure 4. The correspondence of the GSGM model (Lerdahl & Jackendoff, 1983) with participant responses in Experiment 2. "O" refers to an overall score collapsed across GSGM rules. "R" and "A" refer to rest and attack proximity rules. "Reg," "Dyn," "Art," "Len," and "Tim" refer to GSGM register, dynamics, articulation, length, and timbre similarity rules. The horizontal gray bar reflects the range of chance performance. Error bars reflect standard error of the mean.

points of n_3 and n_4 is greater than that between n_1 and n_2 and between n_2 and n_3, then the transition from n_3 to n_4 may be heard as a group boundary" (p. 537). Such a mechanism could account for the observed differences between long-short transitions and short-long transitions and is consistent with the PPS model gap rule. As shown in Figure 6, if the articulation rule is used and notes go from long to short, the gap score is larger at the plus-one boundary. However, if the length rule is used and notes go from long to short, the gap score is larger at the exact boundary. Conversely, with the short-long transitions, if the articulation rule is used, the gap score is larger at the exact boundary. Similarly, if the length rule is used, the gap score is larger at the plus-one boundary. Thus, parsimony suggests combining rest, attack, length, and articulation rules into a single rule similar to the PPS gap rule, with larger gaps more strongly suggestive of phrase boundaries.

Experiment 3: Testing Common Stimuli

Experiments 1 and 2 offered experimental assessments of response/model correspondence of the PPS and GSGM models. In those experiments, each model was tested on stimuli tailored specifically to that model in order to maximize the likelihood of finding higher correspondence between the predicted theoretical segmentation points and the judgments of segmentation by participants; as such, those experiments represented best case scenarios for the corresponding models. Results from Experiments 1 and 2 suggested that the PPS model gap rule is a driving force in the perception of musical segmentation. However, in order to more directly compare the PPS and GSGM models, it is important to test these two models using both a common (i.e., shared) set of stimuli and musical stimuli drawn from the literature that are likely to be more generalizable. Accordingly, the correspondence of predicted segmentation points for the PPS model and for the GSGM model to participants' perceived segmentations on musical stimuli drawn from the sight-singing repertoire and common practice period was examined in Experiment 3.

Method

Participants. Twenty-eight participants were recruited from the same participant pool used in Experiments 1 and 2. None had participated in the previous experiments. The data from five participants were dropped from subsequent analysis for failure to follow directions (they all drew multiple segmentation marks per stimulus). One additional participant did not attend the second session. This resulted in a total of 22 participants (6 male, 16 female) with usable data for the analyses. (For information regarding musical experience of the participants, see section on *Preliminary analysis*, below.)

Apparatus. The computing/headphone apparatus was the same as in Experiments 1 and 2.

Stimuli and response materials.

Auditory stimuli. One hundred thirty-three stimuli, 6 to 24 notes in length, were excerpted from Ottman's (2004) *Music for Sight-Singing*, a compilation of primarily monophonic folk songs and common-practice period music (i.e., music from roughly 1600–1900 A.D., subsuming the Baroque, Classical, and Romantic musical eras; Winold, 1975). Use of stimuli drawn from Ott-

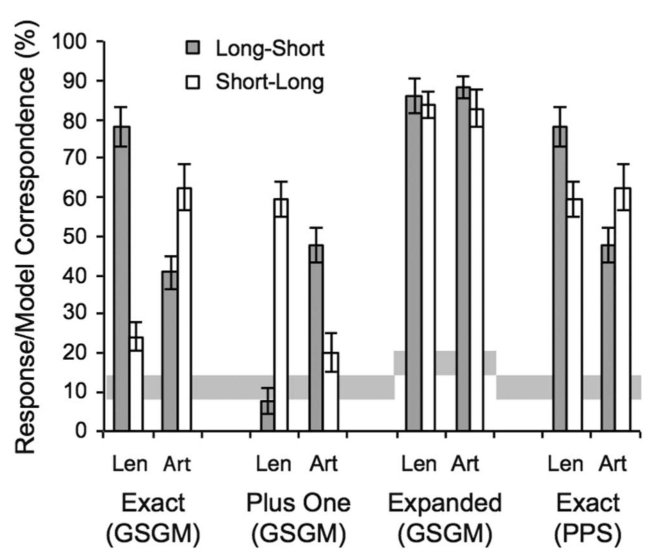

Figure 5. The correspondence of PPS (Temperley, 2001) and GSGM (Lerdahl & Jackendoff, 1983) models with participant responses in Experiment 2. "Len" and "Art" refer to the GSGM length and articulation rules, respectively. "Exact GSGM," "Plus-One GSGM," and "Expanded GSGM" refer to different scoring methods based on the GSGM model and "Exact PPS" refers to scoring based on the PPS model (see text for details on scoring methods). Responses for long-short stimuli are shown by filled (black) columns, and responses for short-long stimuli are shown by unfilled (white) columns. The horizontal gray bar reflects the range of chance performance. Error bars reflect standard error of the mean.

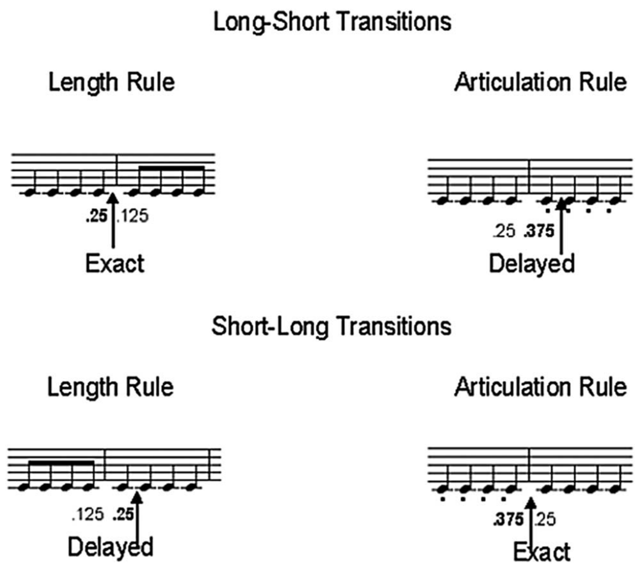

Figure 6. Examples of how the gap rule might drive delayed segmentation in the presence of length changes or articulation changes. Shown below the musical staves are the gap scores for both exact and delayed scoring, with the arrows indicating the theoretical segmentation points based on the gap rule (and the greatest local gap score, as indicated in bold type).

man's book partially addresses the observation of Bruderer et al. (2012) that most studies of musical segmentation involve classical music, as the music in the Ottman compilation involves a broader range of musical styles. In order to control for potential performance issues, auditory presentations of the musical excerpts were generated on the computer using the same software and procedure as in Experiments 1 and 2. All stimuli were evaluated within the framework of both the PPS and GSGM models. Additionally, Ottman's phrase markings are considered to constitute a musically appropriate, professional composer interpretation of phrase boundaries and thus served as an additional point of comparison.

In order to ensure that we were properly assessing both the PPS and GSGM models, we needed to select musical stimuli that would test the rules of both models. The PPS model is more generalizable in that the factors of the model are typically present in all auditory stimuli; it was not difficult to screen the stimuli for these factors. The GSGM model is sufficiently specific that a particular stimulus might or might not contain an opportunity for the test of a given segmentation rule, and thus we took great care to ensure the opportunity to evaluate this model. Therefore, stimulus selection was based on both the PPS and GSGM models.

Stimulus selection based on the PPS model. Gap scores were created for each of the stimuli. Phrase length was carefully monitored such that the number of early, middle, and late theoretical phrase boundaries (and different phrase lengths) were equally represented among the stimuli; theoretical phrase boundaries occurred anywhere from the 3rd to the 18th note of the stimulus. Although the majority of musical examples in Ottman's (2004) compilation were created using parallel construction, an effort was made to ensure that stimuli sampled for Experiment 3 were more representative: 50 were not parallel, 82 were parallel, and 1 phrase boundary was aurally ambiguous (i.e., although technically parallel, the number of notes and rhythm of presentation would not make it clear to a listener that construction was parallel; cf. Temperley & Bartlette, 2002).

Stimulus selection based on the GSGM model. Nineteen stimuli represented each of the GSGM Proximity and Change rules—that is, rest, attack, register, dynamics, length, and timbre (cf. Table 1). The GSGM articulation rule was underrepresented in Ottman (2004) and was thus not represented in the stimuli of Experiment 3.

Response materials. The response materials were the same as in Experiments 1 and 2.

Procedure. Due to the relative large number of stimuli, experimental trials for each participant were divided into two sessions that were scheduled one week apart. Half of the stimuli were presented in Session 1, and the other half were presented in Session 2.

Results and Discussion

Participants' responses were scored as 1 or 0, reflecting a match or nonmatch, respectively, with the theoretical segmentation point derived from the PPS model, and a similar scoring of responses was carried out regarding a match or nonmatch of participants' responses with the theoretical segmentation point derived from the GSGM model. As in Experiments 1 and 2, data were analyzed using GLMMs incorporating random effects using the PROC NLMIXED (i.e., nonlinear, mixed design) procedure in the SAS programming package.

Preliminary analysis. Using the same criteria as in Experiments 1 and 2, participants were classified as low experience (n = 11), intermediate experience (n = 6), or high experience (n = 4). Additionally, one participant did not specify experience information on the questionnaire, and those data were therefore not included in the present analysis. Data were also analyzed using a two-way experience classification (inexperienced, n = 11; experienced, n = 10). In both analyses, there was an effect of low musical experience that did not significantly affect the overall pattern or strength of results; therefore, only data from the three-way (low, intermediate, high) classification are presented in subsequent analyses.

Theoretical segmentation points. The PPS and GSGM models corresponded on theoretical segmentation points for 70.18% of the stimuli; agreement between each of these models and Ottman's (2004) markings were similar, 72.18% and 73.68% agreement, respectively. Although all three sources were in agreement for the majority of stimuli, they were not overly redundant in their predictions (i.e., even though similar total agreement occurred between the models, there were differences in predictions across specific exemplars). Thus, an across-the-board comparison of response/model correspondences could offer considerable insight into the understanding of musical segmentation.

Segmentation responses. A comparison of the PPS and GSGM analyses is shown in Figure 7. Neither model fared particularly well, with the PPS model achieving 53.14% response/model correspondence and the GSGM model achieving 56.42% response/model correspondence. Ottman's markings fared similarly, achieving 53.66% response/model correspondence. Although exceeding chance levels of 4.2% to 16.7%, response/model

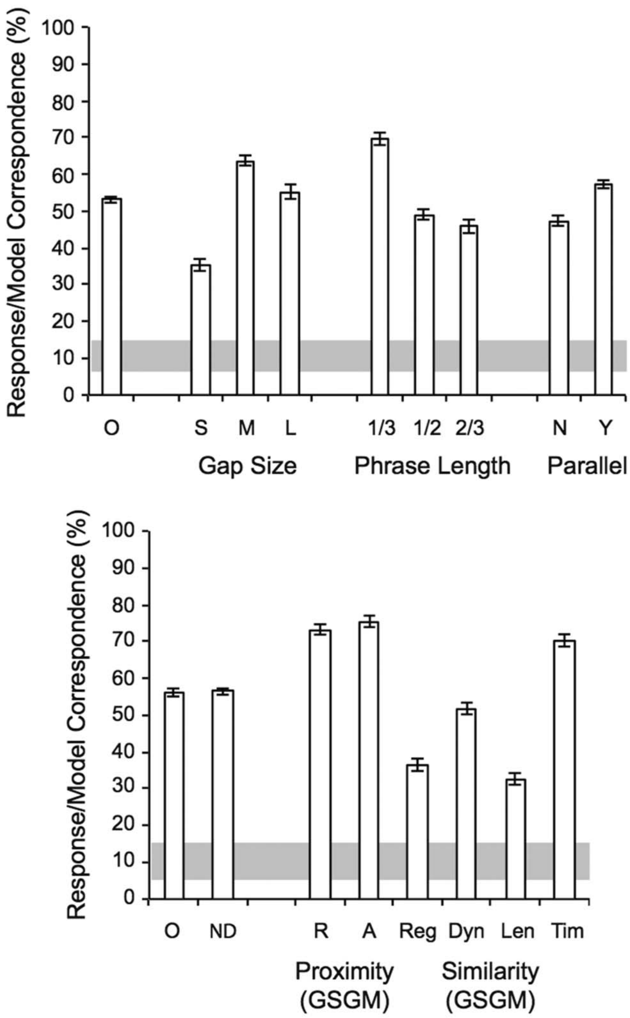

Figure 7. The correspondence of the PPS model (Temperley, 2001; top panel) and the GSGM model (Lerdahl & Jackendoff, 1983; bottom panel) with participant responses in Experiment 3. In the top panel, "O" refers to an overall score collapsed across gap size, phrase length, and parallel structure. "S," "M," and "L" refer to small, medium, and large gap sizes. "1/3," "1/2," and "2/3" refer to whether segmentation was predicted 1/3, 1/2, or 2/3 through the stimulus. "N" and "Y" refer to whether parallelism was not present (i.e., "no") or was present (i.e., "yes"). In the bottom panel, "O" refers to an overall score collapsed across GSGM model rules. "R" and "A" refer to rest and attack proximity rules. "Reg," "Dyn," "Len," and "Tim" refer to register, dynamics, length, and timbre similarity rules. In both panels, the horizontal gray bars reflect the range of chance performance and error bars reflect standard error of the mean.

correspondence was relatively poor compared with that of Experiments 1 and 2. The difference in levels of response/model correspondence across experiments is likely related to the type of stimuli used across the different experiments. Namely, the natural musical stimuli in Experiment 3 were more complex, and likely less predictable, than the musical stimuli in Experiments 1 and 2 that were actually constructed to maximize response/model correspondence. Indeed, Shaefer, Murre, and Bod (2004) suggest that reliance on grouping laws, such as the ones underlying the PPS and GSGM models, is rather limited. It might be that rules are not differentiated sufficiently to account for ambiguities that occur in a more complex musical context. Another possibility is that a broader context is needed in order to elicit all of the rules (i.e., the whole composition, rather than just a brief excerpt, is needed in order to elicit effects of the rules; cf. Bruderer et al., 2012). This suggests that the hierarchical conception of the GSGM model might be critical in musical perception and understanding. Music is generally perceived in the context of a complete song or musical piece, and this wider context evoked by the hierarchical organization of GSGM creates a backdrop in which the full breadth of rules, such as those of the GSGM model, are more likely to be applied.

The PPS model. Analysis of the PPS model yielded positive relationships between response/model correspondence and medium gap size ($\beta = .86$, $p < .0001$), short phrase length ($\beta = 1.55$, $p < .0001$), and medium phrase length ($\beta = .31$, $p = .0094$). Nonparallel construction ($\beta = .49$, $p = .0009$) increased response/model correspondence, whereas small gap size ($\beta = -.92$, $p < .0001$), low musical experience ($\beta = -.78$, $p = .0007$), nonparallel construction at medium gaps ($\beta = -1.65$, $p < .0001$), nonparallel construction with short phrase lengths ($\beta = -.63$, $p = .0057$), and medium phrase lengths at small gaps ($\beta = -.47$, $p = .0229$) decreased response/model correspondence. That is, response/model correspondence increased if gaps were of medium size, phrases were of short or medium duration, and construction was not parallel. Conversely, response/model correspondence decreased if gaps were of small size, participants were low in musical experience, nonparallel construction occurred at medium gaps or with short phrase lengths, and medium phrase lengths occurred with small gaps. Count R-squared was .70.

In an analysis of the PPS model in Experiment 3, and in contrast to the analysis in Experiment 1, all three theoretical factors (gap size, phrase length, parallel construction) were important in determining segmentation responses. Additionally, response/model correspondence decreased in participants with less musical experience. This latter finding suggests there may be a learning component to local context segmentation. Indeed, both the PPS model and the GSGM model were created to reflect the segmentation typically used by musically experienced listeners. Although some researchers suggest that at least local-context segmentation rules apply to the population at large (Deliège, 1987), the results of Experiment 3 are consistent with the idea that musical experience might be an important factor in determining where listeners place phrase boundaries in musical segmentation.

The GSGM model. In an analysis of the GSGM model, rest ($\beta = 1.01$, $p < .0001$), attack ($\beta = 1.16$, $p < .0001$), and timbre changes ($\beta = .87$, $p < .0001$), were positively related to response/model correspondence, whereas register ($\beta = -.72$, $p < .0001$), length ($\beta = -.90$, $p < .0001$), and a low level of musical experience ($\beta = -1.10$, $p < .0001$), were negatively related to response/model correspondence. As noted earlier, stimuli based on the GSGM articulation rule were not available in Experiment 3. Count R-squared was .70.

If expanded scoring (see Experiment 2) was used in evaluating the GSGM model, the negative relationship between length and response/model correspondence ($\beta = -.90$, $p < .0001$) became a positive relationship ($\beta = .54$, $p < .0001$). However, the relationships of other variables to response/model correspondence retained their strength and direction if expanded scoring was used. This result is consistent with use of the gap rule (i.e., as in the short-long transition and long-short transition analyses of Experiment 2). Because articulation was not used as a variable in Experiment 3, the exact data pattern cannot be examined as clearly as in Experiment 2. However, with the application of the gap rule, length was positively related to response/model correspondence. More broadly, as seen in Figure 7, response/model correspondence was higher for the GSGM model rules of rest and attack than for the PPS gap rule. Rests and attacks are generally the largest instances of when a gap occurs, however.

Summary of Experiment 3 analyses. Both the PPS and the GSGM models performed at greater than chance levels in predicting segmentation responses. The degree to which the models did not align with segmentation responses, however, suggested that the hierarchical conception of the GSGM model is likely important for musical perception and understanding given a full musical—not merely local-context—backdrop, and the PPS model gap size concept was shown to explain some of the nuances within the GSGM framework. Neither model was complete in explaining segmentation responses.

General Discussion

The accuracy of the PPS model of musical segmentation (Temperley, 2001; see Table 1) was similar to the accuracy of the GSGM model of musical segmentation (Lerdahl & Jackendoff, 1983; see Table 1) in predicting the perceived segmentation points of musically experienced/trained participants and musically inexperienced/untrained participants. In Experiment 1, the PPS model accounted for 88.66% of segmentation responses, although gap size was the only significant factor. In Experiment 2, GSGM local context segmentation rules accounted for 69.49% of segmentation responses. Consistent with Deliège (1987) and with Frankland and Cohen (2004), an analysis of long-to-short transitions and short-to-long transitions suggested that several GSGM rules could be parsimoniously combined to be functionally similar to the PPS gap rule. In Experiment 3, which used a set of common stimuli derived from the sight-singing repertoire (i.e., Ottman, 2004), neither the PPS model nor the GSGM model stood out as significantly more accurate or more complete. However, results from Experiment 3 suggested that the GSGM length rule could be subsumed under the rubric of the PPS gap rule.

Our findings suggest that musical segmentation can be highly predictable using local context rules. For example, in Experiment 1, 88.66% of responses were accounted for solely by the PPS gap rule. Indeed, the results of Experiment 1 considered in isolation suggest that the GSGM rest, attack point, length, and articulation rules can all be combined into a single quantifiable gap score. Interestingly, such a restructuring would bring musical grouping preference rules more in line with linguistic theory (but see Lerdahl & Jackendoff, 1983): Friederici and Alter (2004) described what they call intonational pauses, or intonational phrase boundaries, in which there is either a pause or a word that is prolonged slightly such that the appropriate syntactic phrase boundaries are conveyed to the listener. In written form, this would be somewhat like including the use of commas to set off an idea or a phrase. A gap might function similarly in a musical context, giving the listener a strong proximity cue (cf. Bruderer et al., 2012).

Given that the models evaluated here performed equally well on actual musical stimuli, we might expect the PPS gap rule to be the principal factor driving segmentation because of implied parsimony, whereas other factors might account for the remaining variance. However, two caveats should accompany any conclusion that the PPS gap rule (or temporal proximity more generally) offers a complete answer to musical segmentation. First, data from Experiment 3 illustrated that more natural musical stimuli are somewhat ambiguous; the models (and composer renderings in Ottman, 2004) accounted for only ~50% of segmentation responses. These results suggest that there are other factors that contribute to musical segmentation (e.g., tonality; see Tan, Aiello, & Bever, 1981). Second, the overarching factor structure suggested by the GSGM model is different from the factor structure implied by the PPS gap rule (see Table 1). To address this issue, future research should incorporate singular value decomposition methods, such as factor analysis, in order to determine more definitively the relationship between segmentation rules and to evaluate factors that could have led to nonpredicted segmentations.

One of the purposes of the experiments reported here was to examine how well the PSS and GSMS models could account for musical segmentation in musically inexperienced and untrained listeners as well as in musically experienced and trained listeners. However, the number of participants classified as low experience, intermediate experience, or high experience varied widely across experiments. These differences might initially appear of concern, but they do not challenge the findings reported here for at least three reasons. First, although in Experiments 1 and 2, model performances were compared across different participant groups, in Experiment 3, model performances were compared within a single participant group. Similarities in these comparisons suggest that participant groups did not significantly differ across experiments. Second, if effects of musical experience were being skewed by extreme values, three-way classifications of musical experience in Experiments 1 and 3 would have yielded different results than did the two-way classifications of musical experience. Third, it is of interest whether either the PPS model or the GSGM model is capable of accounting for musical segmentation of participants who are not musically trained. Along these lines, participants in Experiments 1, 2, and 3 reflect the range of musical experience found in a more general (albeit primarily university-based) population, in which participants might have had some experience (e.g., playing in a school band), but do not necessarily consider themselves musically trained (see discussion in Experiment 1). Indeed, such a range of experience levels offers useful tests of the models and provides greater generalizability of our findings.

Including musically untrained participants in the study was central to our questions of interest, and yet offered methodological challenges. Such participants cannot notate on a musical staff what is heard and mark the phraseology accordingly. However, such individuals can mark a line in a row of dots in short sequences such as those used in Experiments 1–3 (also see Deliège, 1987) or they can press a button when listening to longer sequences (cf. Frankland & Cohen, 2004). We elected to use the dot-and-line-drawing method because we were primarily interested in beginning our investigation with precise, local-context level segmentations; pressing a button occurs in real-time and delays might be involved regarding boundary placements. Also, musically untrained participants might not be as familiar with or sensitive to higher-order or hierarchical rules that are presumably involved in longer sequences. Nonetheless, future studies should continue to investigate models of hierarchical segmentation in the context of longer musical excerpts (cf. Frankland & Cohen). This would allow for the testing of the more hierarchical rules of the full GSGM model and also for testing the PPS model with phrases that would vary from the preferred eight-note length used in the current study. Testing the PPS model on longer phrases is of particular importance, as Temperley (2001) suggested the likelihood of segmentation with such longer phrases would be adjusted by $|(\log_2 N)\text{-}3|$, with N representing the number of notes in the phrase. Such an adjustment is more appropriate in a hierarchical musical setting rather than in the local-context setting of the current experiments. In the context of such an adjustment, an eight-note phrase might not always be preferred.

In summary, the present findings suggest that PPS gap size is the primary factor of musical segmentation. The PPS gap rule has two significant benefits: First, it is universally applicable to music. Gap size can be calculated for every musical stimulus. Second, it is implicitly based on magnitude such that bigger gaps garner more attention than do smaller gaps. If music is straightforward and simple, as in experiments reported here, gaps appear to be the main force behind musical segmentation decisions. Lerdahl and Jackendoff (1983) alluded to this idea when they discussed the GSGM rest and attack local context rules as being primary and other local context rules as being secondary. However, if gaps are small, or ambiguity is created in some other fashion (a goal that can be achieved in multiple ways, depending on the composer's preferences), other factors such as phrase length and parallelism (Temperley, 2001) or register and timbre (Lerdahl & Jackendoff, 1983) become more salient. Even so, gap size appears to provide fundamental information regarding the structure of musical phrases, and this information is used by musically experienced and musically inexperienced listeners. Thus, gap size appears to be the predominant cue to musical segmentation.

References

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Hoboken, NJ: Wiley and Sons. http://dx.doi.org/10.1002/0471249688

Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature, 19,* 1483–1536.

Bartlett, J. C., & Dowling, W. J. (1980). Recognition of transposed melodies: A key-distance effect in developmental perspective. *Journal of Experimental Psychology: Human Perception and Performance, 6,* 501–515. http://dx.doi.org/10.1037/0096-1523.6.3.501

Besson, M., & Friederici, A. D. (1998). Language and music: A comparative view. *Music Perception, 16,* 1–9. http://dx.doi.org/10.2307/40285773

Blacking, J. (1973). *How musical is man?* Seattle, WA: University of Washington Press.

Brent, M. R. (1999). An efficient, probabilistically sound algorithm for segmentation and word discovery. *Machine Learning, 34,* 71–105. http://dx.doi.org/10.1023/A:1007541817488

Bruderer, M. J., McKinney, M. F., & Kohlrausch, A. (2012). Perceptual evaluation of musicological cues for automatic song segmentation. *Psychomusicology: Music, Mind, and Brain, 22,* 3–17. http://dx.doi.org/10.1037/a0026872

Cambouropoulos, E. (2001). The local boundary detection model (LBDM) and its application in the study of expressive timing. In *Proceedings of the International Computer Music Conference (ICMC2001),* San Francisco, CA: International Computer Music Association.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1968). *Language and mind*. New York, NY: Harcourt Brace Jovanovich. http://dx.doi.org/10.1037/e400082009-004

Chomsky, N. (2002). *Syntactic structures* (2nd ed.). New York, NY: Mouton de Gruyter. (Original work published 1957) http://dx.doi.org/10.1515/9783110218329

Deliège, I. (1987). Grouping conditions in listening to music: An approach to Lerdahl & Jackendoff's grouping preference rules. *Music Perception, 4,* 325–359. http://dx.doi.org/10.2307/40285378

Deutsch, D., & Feroe, J. (1981). The internal representation of pitch sequences in tonal music. *Psychological Review, 88,* 503–522. http://dx.doi.org/10.1037/0033-295X.88.6.503

Dowling, W. J. (1984). Musical experience and tonal scales in the recognition of octave scrambled melodies. *Psychomusicology: A Journal of Research in Music Cognition, 4,* 13–32. http://dx.doi.org/10.1037/h0094206

Drake, C. (1998). Psychological processes involved in the temporal organization of complex auditory sequences: Universal and acquired processes. *Music Perception, 16,* 11–26. http://dx.doi.org/10.2307/40285774

Frankland, B. W., & Cohen, A. J. (2004). Parsing of melody: Quantification and testing of the local grouping rules of Lerdahl and Jackendoff's A Generative Theory of Tonal Music. *Music Perception, 21,* 499–543. http://dx.doi.org/10.1525/mp.2004.21.4.499

Friederici, A. D., & Alter, K. (2004). Lateralization of auditory language functions: A dynamic dual pathway model. *Brain and Language, 89,* 267–276. http://dx.doi.org/10.1016/S0093-934X(03)00351-1

Gregory, A. H. (1978). Perception of clicks in music. *Perception and Psychophysics, 24,* 171–174. http://dx.doi.org/10.3758/BF03199545

Jackendoff, R., & Lerdahl, F. (1981). Generative music theory and its relation to psychology. *Journal of Music Theory, 25,* 45–90. http://dx.doi.org/10.2307/843466

Jescheniak, J. D., Hahne, A., & Friederici, A. D. (1998). Brain activity patterns suggest prosodic influences on syntactic parsing in the comprehension of spoken sentences. *Music Perception, 16,* 55–62. http://dx.doi.org/10.2307/40285777

Johnson, N. F. (1965). The psychological reality of phrase-structure rules. *Journal of Verbal Learning and Verbal Behavior, 4,* 469–475. http://dx.doi.org/10.1016/S0022-5371(65)80044-5

Krumhansl, C. L. (1998). Topic in music: An empirical study of memorability, openness, and emotion in Mozart's String Quintet in C Major and Beethoven's string quartet in A minor. *Music Perception, 16,* 119–134. http://dx.doi.org/10.2307/40285781

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Long, J. S., & Freese, J. (2006). *Regression models for categorical outcomes using stata* (2nd ed.). College Station, TX: Stata Press.

Ottman, R. W. (2004). *Music for sight singing* (6th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Patel, A. D. (1998). Syntactic processing in language and music: Different cognitive operations, similar neural resources? *Music Perception, 16,* 27–42. http://dx.doi.org/10.2307/40285775

Patel, A. D. (2010). *Music, language, and the brain*. New York, NY: Oxford University Press.

Pearce, M. T., Müllensiefen, D., & Wiggins, G. A. (2010). Melodic grouping in music information retrieval: New methods and applications. In Z. W. Ras & A. Wieczorkowska (Eds.), *Advances in music information retrieval* (pp. 364–388). Berlin: Springer. http://dx.doi.org/10.1007/978-3-642-11674-2_16

Peretz, I. (1989). Clustering in music: An appraisal of task factors. *International Journal of Psychology, 24,* 157–178. http://dx.doi.org/10.1080/00207594.1989.10600040

Peretz, I. (1993). Auditory atonalia for melodies. *Cognitive Neuropsychology, 10,* 21–56. http://dx.doi.org/10.1080/02643299308253455

Peretz, I., Kolinsky, R., Tramo, M., Labrecque, R., Hublet, C., Demeurisse, G., & Belleville, S. (1994). Functional dissociations following bilateral lesions of auditory cortex. *Brain: A Journal of Neurology, 117,* 1283–1301. http://dx.doi.org/10.1093/brain/117.6.1283

Raffman, D. (1993). *Language, music, and mind*. Cambridge, MA: MIT Press.

Rosenberg, S. (1968). Association and phrase structure in sentence recall. *Journal of Verbal Learning and Verbal Behavior, 7,* 1077–1081. http://dx.doi.org/10.1016/S0022-5371(68)80071-4

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274,* 1926–1928. http://dx.doi.org/10.1126/science.274.5294.1926

SAS Institute. (2003). *PROC NLMIXED compared with other SAS procedures and macros*. Retrieved from http://support.sas.com/91doc/getDoc/statug.hlp/nlmixed_sect4.htm#topofpage. Cary, NC.

Shaefer, R. S., Murre, J. M. J., & Bod, R. (2004). Limits to universality in segmentation of simple melodies. In S. D. Lipscomb, R. Ashley, R. O. Gjerdingen, & P. Webster (Eds.), *Proceedings of the 8th International Conference on Music Perception & Cognition* (pp. 247–250). Adelaide, Australia: Causal Productions.

Sloboda, J. (1985). *The musical mind: The cognitive psychology of music*. Oxford: Clarendon Press.

Swinney, D., & Love, T. (1998). The processing of discontinuous dependencies in language and music. *Music Perception, 16,* 63–78. http://dx.doi.org/10.2307/40285778

Tan, N., Aiello, R., & Bever, T. G. (1981). Harmonic structure as a determinant of melodic organization. *Memory and Cognition, 9,* 533–539. http://dx.doi.org/10.3758/BF03202347

Temperley, D. (2001). *The cognition of basic musical structures*. Cambridge, MA: MIT Press.

Temperley, D., & Bartlette, C. (2002). Parallelism as a factor in metrical analysis. *Music Perception, 20,* 117–149. http://dx.doi.org/10.1525/mp.2002.20.2.117

Tenney, J., & Polansky, L. (1980). Temporal Gestalt perception in music. *Journal of Music Theory, 24,* 205–241. http://dx.doi.org/10.2307/843503

Verheul, M. H. G., & Geuze, R. H. (2004). Bimanual coordination and musical experience: The role of intrinsic dynamics and behavioral information. *Motor Control, 8,* 270–291.

Wallin, N. L., Merker, B., & Brown, S. (Eds.). (2000). *The origins of music*. Cambridge, MA: MIT Press.

Windmeijer, F. A. G. (1995). Goodness-of-fit measures in binary choice models. *Econometric Reviews, 14,* 101–116. http://dx.doi.org/10.1080/07474939508800306

Winograd, T. (1968). Linguistics and the computer analysis of tonal harmony. *Journal of Music Theory, 12,* 2–49. http://dx.doi.org/10.2307/842885

Winold, A. (1975). Rhythm in twentieth-century music. In G. Wittlich (Ed.), *Aspects of twentieth century music* (pp. 208–269). Englewood Cliffs, NJ: Prentice Hall.

Yngve, V. H. (1960). A model and an hypothesis for language structure. *Proceedings of the American Philosophical Society, 104,* 444–466.

Zbikowski, L. M. (1999). Musical coherence, motive, and categorization. *Music Perception, 17,* 5–42. http://dx.doi.org/10.2307/40285810

Received February 19, 2014
Revision received February 1, 2015
Accepted April 15, 2015 ■