Downbeat delays are a key component of the swing feel in jazz

Corentin Nelias[1,2], Eva Marit Sturm[3], Thorsten Albrecht[3], York Hagmayer[3], and Theo Geisel[1,2,4*]

1 Max Planck Institute for Dynamics and Self-Organization, 37077 Göttingen, Germany
2 Department of Physics, Georg-August-University Göttingen, 37077 Göttingen, Germany
3 Georg-Elias-Müller Institute of Psychology, Georg-August-University Göttingen, 37073 Göttingen, Germany
4 Bernstein Center for Computational Neuroscience, 37077 Göttingen, Germany

* theo.geisel@ds.mpg.de

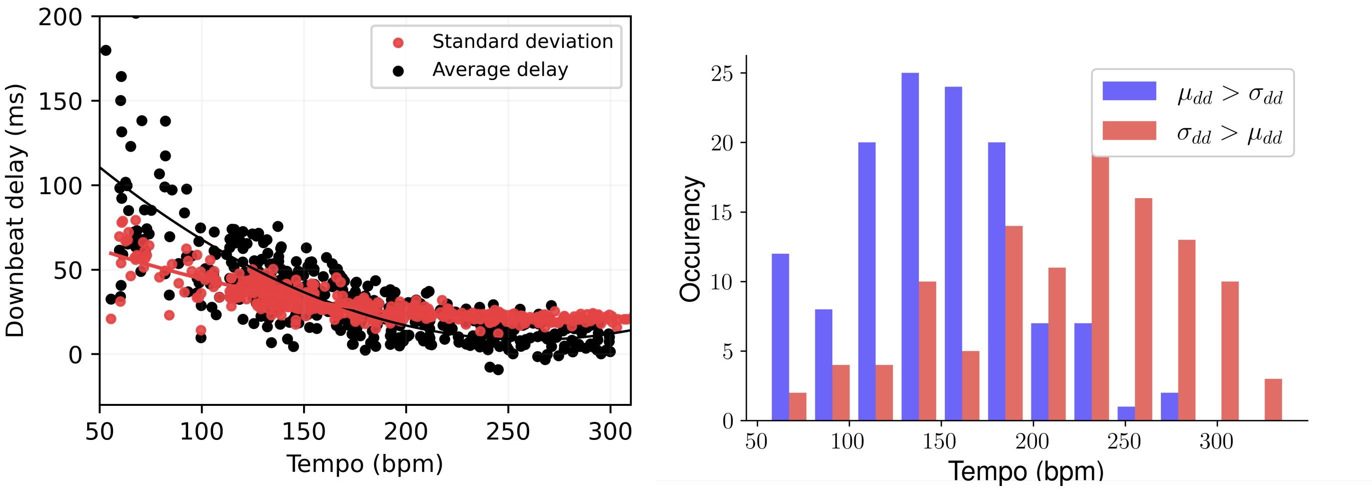
Contents

1 Timing analysis of jazz solos

1.1 Downbeat timing

1.1.1 Dependence on tempo and swing ratio

In the main body of the article, we inferred the behavior of downbeat delays by studying average downbeat placements. For each piece in the Weimar Jazz Database, we computed the mean downbeat position and noticed that downbeats tend to be delayed. Since the average downbeat delay values are typically well above 0, we described them as a form of *systematic* microtiming deviation. However, this description and what we infer from the dependence on tempo of downbeat delays only make sense if the standard deviations of the delays are smaller than their average absolute values. For this reason, we determined the standard deviation of the downbeat placement for each piece, as presented in Fig. 1a. The value of the standard deviation for tempi inferior to 200 bpm lies below the typical downbeat delay (as shown by the black and red fit lines) suggesting that on average, there is always a downbeat delay being played even when the fluctuations are considered. This becomes even more conspicuous in Fig. 1b, which shows the proportion of pieces having an average downbeat delay value higher than its standard deviation. The situation is different in the high tempo range, where the average delay values go below 20 ms, which is close to the perception threshold below which no effect or difference would be detected anymore. With increasing tempi in this high tempo range, the downbeat delays tend to 0 while the standard deviation stabilizes around 25 ms. Thus, the delays for high tempo may be influenced by involuntary fluctuations inherent in the motoric process of musical performance, as they stem from close-to-random microtiming deviations. The tempo is not the only variable influencing



(a) (b)

Figure 1: **Comparison between average downbeat delay and standard devation** as a function of tempo. **a)** Average downbeat delays (in black) as well as the corresponding standard deviation (in red) for all solos of the Weimar Jazz Database. Each dot represents one solo. **b)** Distribution of solos as a function of tempo. Solos having an average value of downbeat delays (μ_{dd}) larger than their standard deviation (σ_{dd}) are shown in blue. Conversely, pieces in which μ_{dd} is smaller than σ_{dd} are shown in red. Non-swinging genres (e.g. "fusion") were excluded. From these two graphs, it is apparent that downbeat delays are likely to have positive values for tempi below 200 bpm. Indeed, in this tempo range, most solos present downbeat delays having an average value higher than their standard deviation.

downbeat delays. Indeed, the soloist's swing ratio also seems to play a role in the expression of downbeat delays. Figure 2 shows the average downbeat delay as a function of swing ratio and tempo. We notice that pieces that are not swinging (i.e. with a swing ratio close to 1) present little downbeat delays. This supports the hypothesis that downbeat delays are characteristic of *swing*. The dependency between downbeat delays and swing ratio, however, is not readily discernible from this graph. To clarify this dependency, we applied a principal component analysis (PCA) to the data in Figure 2. The results are shown in Figure 3. After an appropriate transformation[1], we project each point onto the two principal components. This leads to a reduction of the dimensionality of the data from 3 (swing ratio, downbeat delay, tempo) to 2 (swing ratio, downbeat delay). The result is then projected onto the original swing-ratio/downbeat delay plane to allow an interpretation of the results. From Figure 3, one can clearly discern a linear dependence between downbeat delays and mean swing ratio. This dependency is a function of tempo, however: the slope of the linear relationship between downbeat delay and swing ratio is almost entirely determined by the tempo.

1.1.2 Dependence on position within the measure

In jazz as well as in classical music, the concept of *strong* and *weak* beat is used to denote the importance of certain downbeats w.r.t others. In jazz, when a measure has 4 downbeats, downbeats 2 and 4 are referred to as *strong beats*. One may ask whether the position of a downbeat within the measure has an influence on

[1]Each data point is shifted and re-scaled to set the mean and variance of the data to 0 and 1, respectively. This operation is not reversible after the projection of the data onto its principal components.

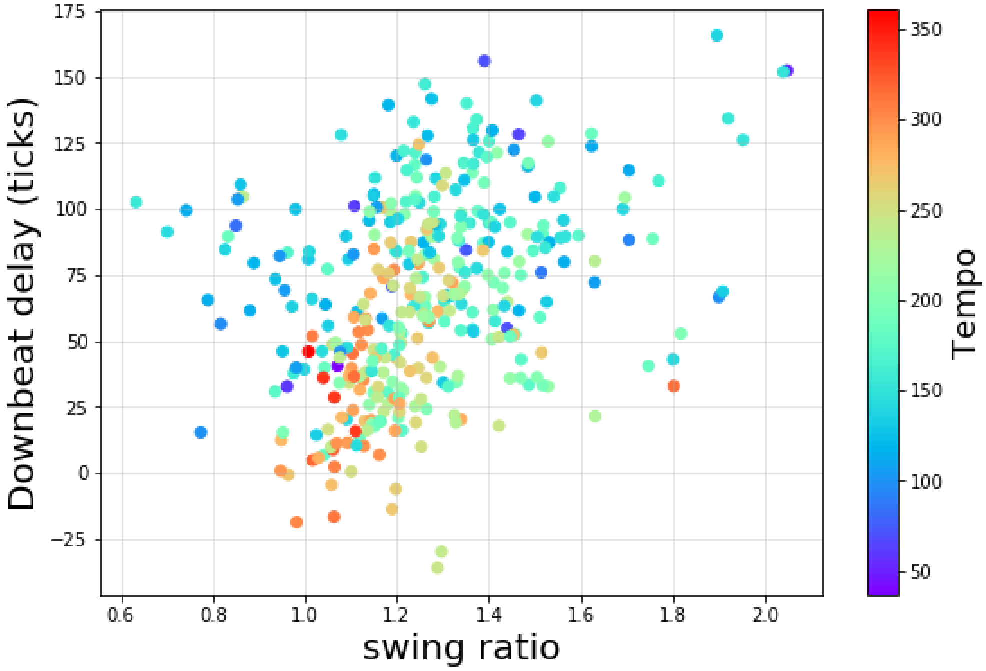


Figure 2: **Average downbeat delays as a function of swing ratio.** The tempo varies between 50 and 350 bpm and is represented by a color coding going from deep blue to red. Pieces with a swing ratio close to 1 (likely not swinging) do not present significant downbeat delays.

its average delay. To answer this question, we plot the average downbeat delays for each downbeat position in the measure. Results are shown in Fig. 4. We do not see any noticeable dependence on the position within the measure. The variation of average downbeat delays as a function of tempo is nearly identical and independent of the downbeat position in the measure. Individual examples such as the ones presented in Fig. 5 also tend to show an absence of dependence of the downbeat delays on the position in the measure. We note that there are some variations among musicians. For some pieces, different quarter note values do have different downbeat delays. However, no systematic behavior is observed and as we already stated, an influence of the position within the measure is on average not noticeable.

1.1.3 Dependence on jazz subgenre

Looking at individual pieces, we noticed some variations of the downbeat delay values (with or without specifying the position in the measure). This variability may be explained by the wide variety of subgenres that are available in the Weimar Jazz Database: it spans entries dating back to 1927 as well as more recent recordings from 2009. For this reason, we also looked at the dependence of downbeat delays on the jazz subgenre. The results are presented in Fig. 6 and show that downbeat delays as a function of tempo behave similarly for jazz subgenres for which *swing* is an essential component (swing, bebop or hardbop). In the case of fusion, a style characterized by the inclusion of elements stemming from rock and hardrock music, the downbeat delays behave differently and do not seem to have a very strong dependence on tempo. This is indicating that the downbeat delays are specific for swing.

1.2 Offbeat timing

As downbeat delays alone are not decisive for a positive effect on the swing feel, we also analyzed the timing of offbeats.

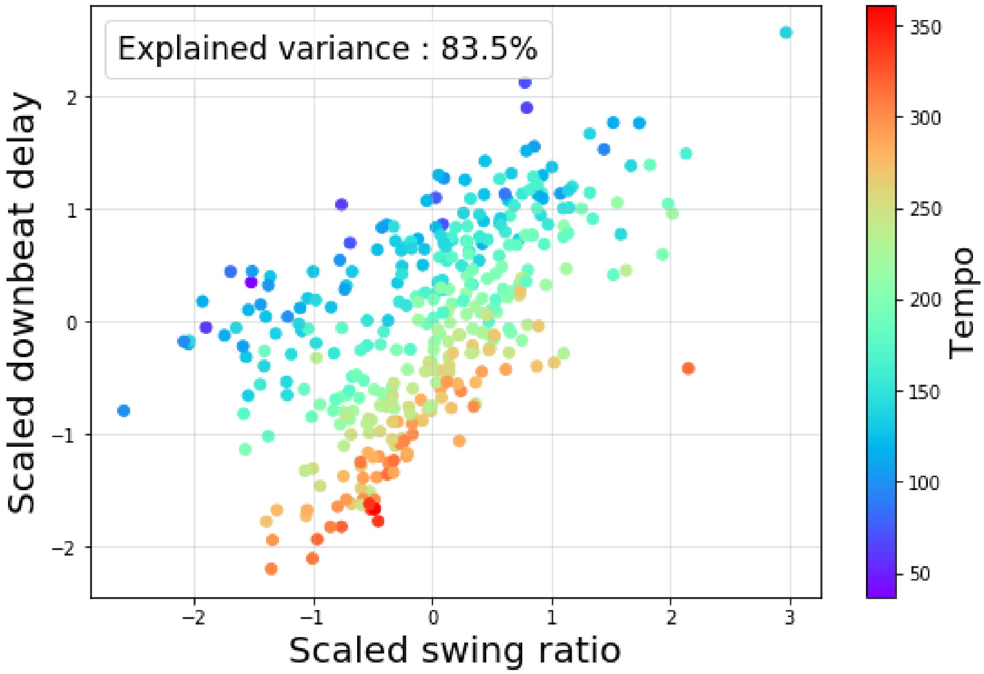


Figure 3: **Principal component analysis of downbeat delays as a function of swing ratio and tempo.** In three dimensions, each piece of the Weimar Jazz Database is represented by a point indicating its average downbeat delay, mean swing ratio, and tempo. This graph shows a projection of all data points onto the first two principal components after proper scaling. There is a clear dependence of downbeat delays on the swing ratio. The relationship between these two variables itself is tempo-dependent.

1.2.1 Dependence on tempo

Without surprise, the variation of the average offbeat position as a function of tempo is similar to that of the swing ratio when considered in ticks. This can be seen in Fig. 7a. An explanation for this is that ticks and swing ratios are both measures of the relative position of offbeats w.r.t the downbeats. If the average offbeat position is plotted in milliseconds (absolute time measurement), this relationship resembles that of the downbeats plotted in milliseconds. The reason for that is that, even if the value in ticks decreases as the tempo decreases, the absolute value of a tick grows with decreasing tempo. Note that it is only possible to measure offbeat positions relative to drum downbeats but not relative to drum offbeats (i.e offbeat deviations), as the Weimar Jazz Database does not report the drum offbeats.

1.3 Triplet timing

1.3.1 Dependence on tempo and position

Downbeat-offbeat note pairs, despite being extremely common in jazz soli, are of course not the only rhythmical values used by improvisers. Another common value, the *triplet*, is constituted of three notes equally spaced (in theory) in the time-span of a quarter note. We could not study the effect of such triplets in the frame of our online experiment because it would have complexified and lengthened the study very much, making it more difficult to find participants. However, we studied the behavior of triplets based on the Weimar Jazz Database. To see whether or not triplets are really played in a regular fashion as is usually assumed, it is helpful to look at the results in ticks. The results are presented in Fig.8, where we plot the average deviation of each individual triplet note w.r.t to its theoretical position. We see that there is a tempo range at which musicians do play regular triplets on average (around 220 bpm). However, for most tempi they do not. In the low tempo range, triplet notes tend to be played further apart from each other than

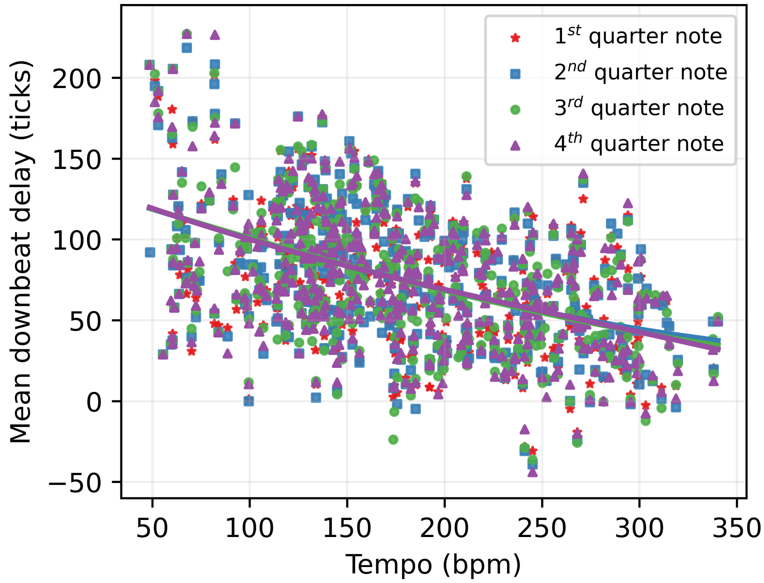
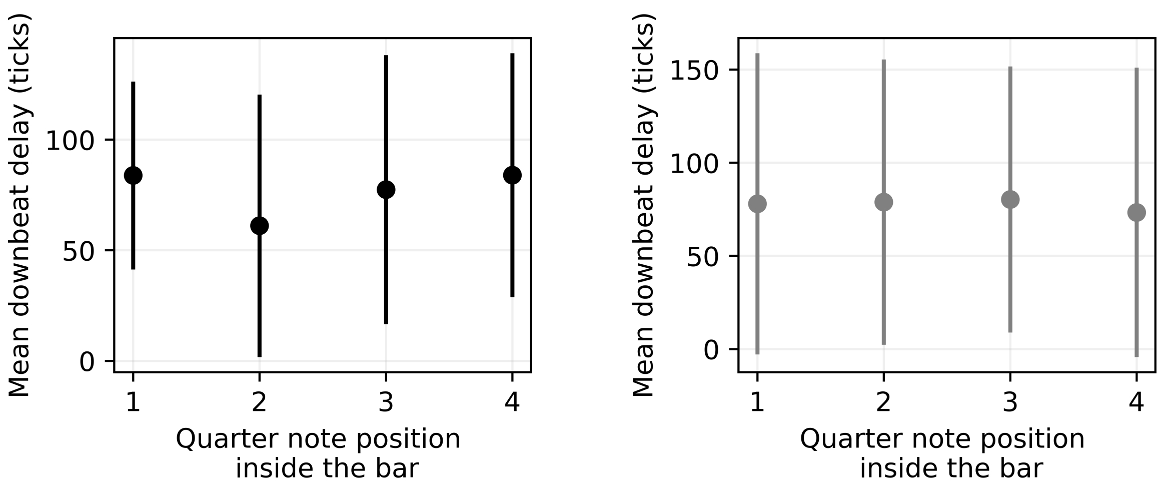


Figure 4: **Average downbeat delay as a function of tempo sorted out by the downbeat position within the measure.** Each color represents a specific beat. The lines are polynomial fits of order 3 for each color and are closely coinciding i.e. not showing a dependence on quarter note position. The delays are presented in ticks to ease comparison.



(a) Benny Carter's "Sweet Lorraine"

(b) Buck Claytons "After Theatre Jump"

Figure 5: **Average downbeat delay (in ms) as a function of quarter note position within the measure.** The two plots pertain to different improvisations by different artists. a) is an improvisation by Benny Carter over "sweet Lorraine" and b) and improvisation by Buck Clayton over "After Theatre Jump". Each point represents the average downbeat delay at the corresponding quarter note position, and the vertical lines show their respective standard deviation.

expected. In the high tempo domain, this relationship is inverted and triplet notes are "contracted", being played closer to each other than regular triplets. The effect of such irregular triplet playing on the swing feel is unclear and could be the subject of further work.

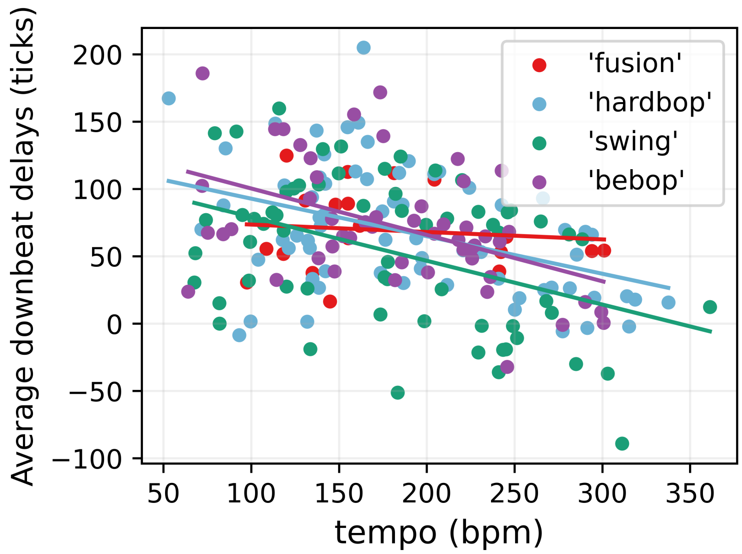


Figure 6: **Average downbeat delay (in ticks) as a function of tempo and jazz subgenre.** Each color represents a different subgenre. The straight colored lines correspond to a linear fit pertaining to genres of the same color code.

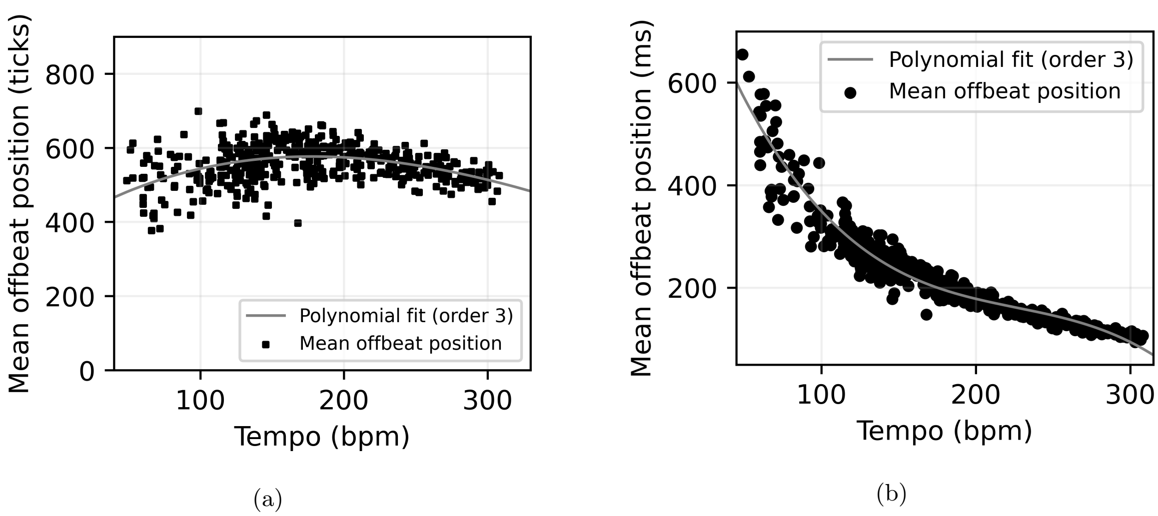


Figure 7: **Average offbeat positions as a function of tempo.** a) in ticks, b) in milliseconds. Offbeat positions are measured relative to the drum downbeats. For example, a value of 600 ticks means 600 ticks after the downbeat. The full lines represent polynomial fits of order 3.

2 Experiment on the perceived swing feel

2.1 Musical background

Musicians participating in the experiment were asked to categorize themselves as professional jazz musicians, semiprofessional jazz musicians, amateur jazz musicians, or non-jazz musicians. We analyzed data only from semiprofessional and professional jazz musicians because only these groups of participants were generally able

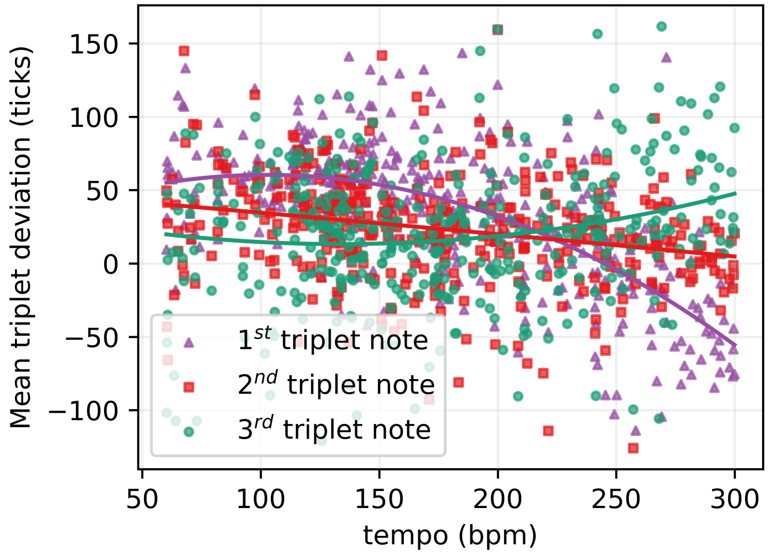


Figure 8: **Average deviation of triplet notes w.r.t. perfectly regular triplets as a function of tempo.** The colors distinguish different positions in a triplet.

	Professional jazz musicians n = 20		Semiprofessional jazz musicians n = 20	
	M	SD	M	SD
Weekly hours of practice	15.75	10.41	9.35	6.12
Concerts played (preceding 12 months)	68.10	53.47	17.80	14.32

Table 1: Sample characteristics of the two groups of jazz musicians participating in the experiment.

to perceive differences reliably[2] between the different versions presented (see Sect. 2.3). Table 1 provides information on weekly hours of practice and concerts played. There was a clear difference between groups, with professional musicians having more experience. Professional musicians practiced more and played significantly more concerts. These differences in experience might explain why professional jazz musicians gave more conservative ratings of the swing feel than semiprofessional jazz musicians (see Fig. 9, which shows histograms of the swing ratings). A considerable number of semiprofessional jazz musicians gave the highest rating to the *downbeat delayed* condition. By contrast, professional jazz musicians were more reluctant to give the highest rating, but more often used the second highest for the *downbeat delayed* condition. Table 2 is reproduced from the main text and shows the results of the statistical analysis of the swing ratings (see Sect. 2.2 for details). There was a significant main effect of musician category with an odds ratio larger than 1 in favor of the semiprofessional jazz musicians. This indicates that semiprofessional jazz musicians gave higher ratings than professional jazz musicians.

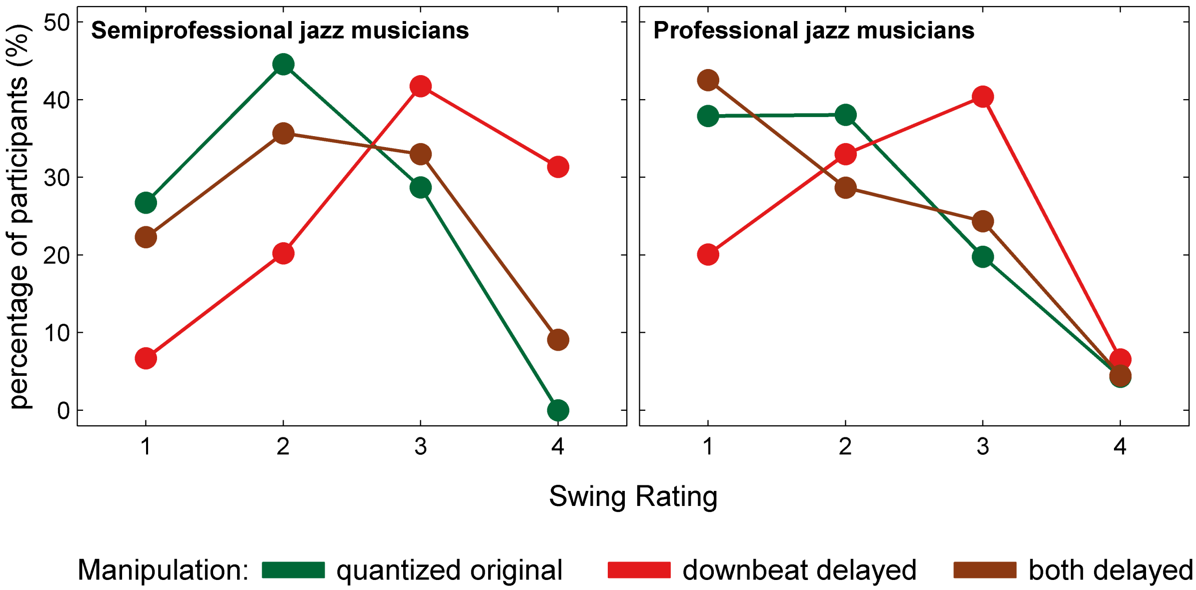


Figure 9: **Percentages of swing ratings for the different manipulations shown separately for each musician category**. The percentages (symbols) of each condition (color coded) add up to 100 percent. The figure shows that the *downbeat delayed* manipulation is - on average - associated with higher swing ratings than the *quantized original* and *both delayed* manipulations.

2.2 Statistical analyses

To analyze how the manipulations affected participants' ratings of the swing feel, we applied an ordinal logistic regression model. Two predictors as well as their interaction were included in the model, the manipulation (*quantized original QO*, *downbeat delayed DD*, and *both delayed BD*) and the musical background (semiprofessional vs. professional jazz musician). The ordinal regression takes into account the ordinal nature of our 4-point rating scale, with ordered but unstructured thresholds for our response categories [1]. Two random effects were included to control for variation in judgments among participants and between pieces. A cumulative link mixed model was fitted with Laplace approximation using the *clmm* function from the ordinal package in *R* [2]. Likelihood-ratio-tests were used to evaluate whether the predictors significantly improved the fit of the model. Formally, the model can be written as

$$\log\left[\frac{\Pr(\,rating \leq j\,)}{1-\Pr(\,rating \leq j\,)}\right] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 \times x_2 \tag{1}$$

Pr gives the probability that the swing rating is j or lower (where j represents the ratings 1 to 4). Variable x_1 represents the manipulation and x_2 the musician category, $x_1 \times x_2$ stands for their interaction. The α_j represent the thresholds in the model. As reference categories, we chose the *quantized original* version and the semiprofessional jazz musicians.

The likelihood-ratio-tests showed that adding manipulation and musician category as predictors to the model improved the model fit, as did the interaction between the two predictors. Therefore we continued using the model as specified formally above. All parameter estimates of the model are presented in Table 2.

Parameter estimates were also transformed into odds ratios as effect sizes. The first two odds ratios in Table 2 indicate how much more likely it is that the respective manipulation elicits higher swing ratings than the

[2]Some amateurs and non-jazz musicians were able nevertheless to perceive differences and expressed preference for the *downbeat delayed* version. We did not have enough participants from these two categories, however, to perform a statistical analysis.

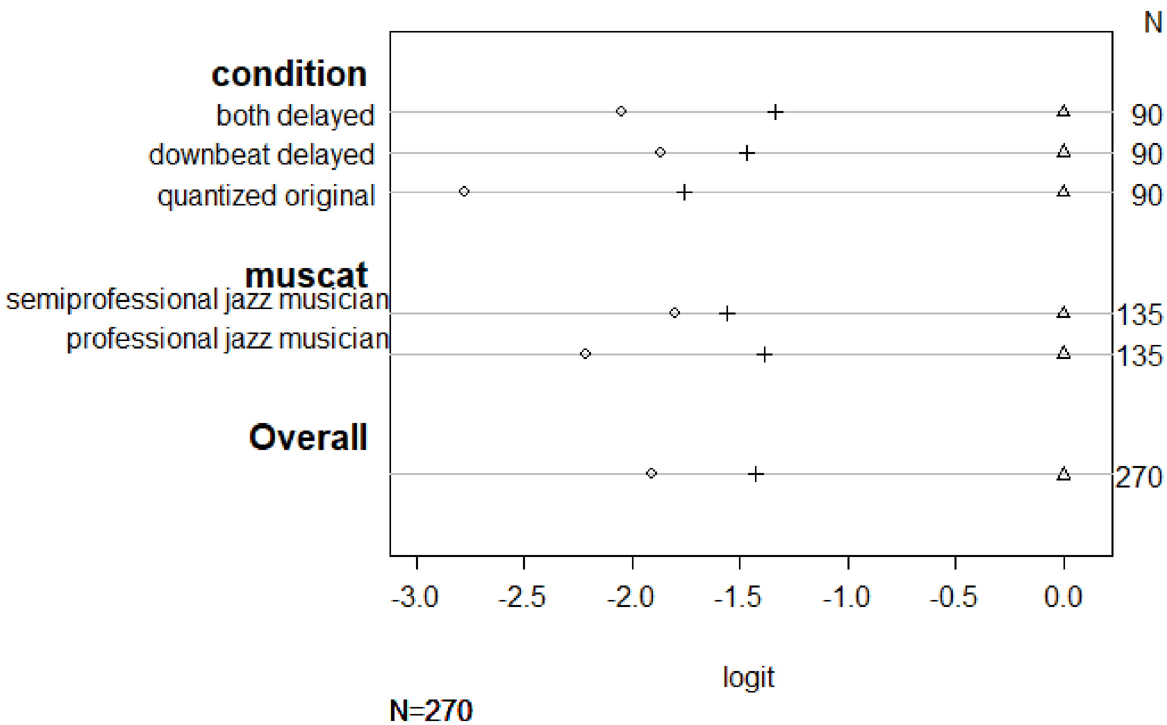


Figure 10: **Visual inspection of the proportional odds assumption.** For our predictors condition (*quantized original*, *downbeat delayed*, *both delayed*) and musician group (muscat: semiprofessional jazz musician, professional jazz musician). Symbols denote the distance in predicted logits between pairs of categories of the dependent variable. logit(Swing Rating ≥ 2) is set as reference point (triangle), logit(swing ≥ 3) − logit(swing ≥ 2) (cross) and logit(swing ≥ 4) − logit(swing ≥ 3) (diamond). N = 270 are the number of data points included in the analysis. If the proportional odds assumption holds, distances between predicted logits for every set of categories of our dependent variable should be equal.The first set of coefficients was normalized to be zero, having a common reference point for easier comparisons. The distances between the set of coefficients for both predictors point in the same direction, but are not equal in size. This indicates that the effects of rating jazz recordings as swinging may not be independent of the levels of the predictors.

Condition	β	SE	p	OR	CI	$Power$
DD vs **QO** (condition 1)	2.012	0.435	$< .001$***	7.48	[3.19 ; 17.54]	1.0
BD vs **QO** (condition 2)	0.263	0.341	.440	1.30	[0.67 ; 2.54]	0.15
semi-pro vs **pro** (muscat)	1.370	0.582	.019*	3.94	[1.26 ; 12.31]	
condition 1 × **muscat**	1.946	0.851	.022*	7.00	[1.32 ; 37.11]	
condition 2 × **muscat**	0.847	0.678	.212	2.33	[0.62 ; 8.82]	
Threshold coefficients				Estimate	Std. error	
1|2				-1.743	0.333	
2|3				0.524	0.315	
3|4				3.632	0.440	

Table 2: **Proportional odds mixed model for swing fitted with Laplace approximation**. The estimates display the effects of the manipulations and musician categories. *** $p < .001$, * $p < .05$. β = regression coefficient. SE = Standard Error. For details on the calculation of statistical power see Sect. 2.2.1. For other abbreviations please refer to Table 1 of the main text.

reference, i.e. the *quantized original* version. A value larger than one signifies a higher probability, a value lower than one would indicate that the probability is reduced.

In an ordinal regression, we have different thresholds for the response categories but constant slopes due to the proportional odds assumption. Using a graphical method, we tested whether the proportional odds assumption did hold for our model [3]. For the different levels of our two predictors, the distances between the predicted logits for each set of categories of the dependent variable should be equal. If the proportional odds assumption holds, the distances between the symbols on each line should be similar for each level of the respective predictor. Looking at our model with the predictors manipulation and musician category, the order of the symbols are equal, but the distances are not the same (Fig. 10). This indicates that the proportional odds assumption is not fully met. Nevertheless, we have no reason to assume that the regression model produced artifacts due to violated assumptions, because the ROC-curves and the regression model showed the same strong significant effects. Moreover, the graphical method does not fully reflect our model, because the two random effects could not be included in the analysis.

2.2.1 Statistical power and robustness

As it is not possible to a priori know or estimate the expected effect size of the two manipulations on the perceived swing feel, we calculated the achieved statistical power in a post hoc analysis. We simulated 1000 data sets using the parameters obtained in Table 2 for the two conditions (DD vs. QO and BD vs. QO), the standard deviations for the random effects, and the actual sample size. Simulations were carried out for a model based on the proportional odds assumption, which included only the two conditions (DD vs, QO and BD vs. QO) while excluding the musician category. This simplification is justified, as we used sum-to-zero contrasts to run our statistical analyses, which entails that the parameters shown in Table 2 give the average effect for the two groups of musicians. Based on the simulations we calculated the relative frequency of data sets, for which each of the two conditions (DD vs. QO and BD vs. QO) reached significance. This relative frequency is an estimate for the statistical power achieved in the study. Note that there are no established statistical packages for power calculations of ordinal regression analyses with random effects. Therefore simulations are the only possible approach to estimate achieved power.

The results of the simulations included in Table 2 are conclusive. The achieved power for the first contrast (DD vs. QO) is 1.0, which means that in each simulated data set this contrast turned out to be significant. This result is not surprising given the very large effect size of $OR = 7.48$ and the within-subjects design of our study. The power of the second contrast (BD vs. QO) is rather low with a power of 0.15. This contrast thus is not significant in most simulated data sets. This is in line with the lack of an effect for the BD vs. QO condition in the findings of our study.

To further evaluate the robustness of our findings, we checked whether there were outliers that could have impacted the estimated effects. To this end, we reiterated the main analysis as described in the previous section 37 times leaving out a different participant in each run. This resulted in 37 ordinal regressions with 36 participants each. If there were biasing outliers, the results would deviate considerably from the ones presented above. It turned out that this was not the case (Fig. 11A). For DD vs. QO, the estimated coefficients range between $\beta = 1.86$ and $\beta = 2.20$ (whole sample $\beta = 2.01$). Two values of β deviate more than 2 standard deviations from the whole sample effect and indicate two participants which have a somewhat larger impact on the effect in the whole sample. But as the two β values lie *above* the upper bound, leaving out these outliers even leads to an *increased* effect in the remaining sample ($\beta = 2.37$), i.e. these two participants show a substantially smaller effect than the whole sample effect. For BD vs. QO, the estimated coefficients range between $\beta = 0.13$ and $\beta = 0.39$, (whole sample $\beta = 0.26$). Three coefficients deviate more than 2 standard deviations from the whole sample effect (two below the lower bound, one above the upper bound) indicating three participants as possible outliers. Excluding these participants leads to an even *weaker* effect for BD vs. QO in the remaining sample ($\beta = 0.14$).

Finally, we carried out another check for the robustness of our effects by reiterating the ordinal regressions with increasingly smaller sample sizes down to n = 13. To this end, we randomly removed one to six participants from each category before computing the ordinal regression as described above. For each sample size we computed 50 regressions. The resulting coefficients for DD vs. QO and BD vs. QO are shown in Fig. 11B. For DD vs. QO the effect remains stable and well above zero down to a sample size of 25 participants. For BD vs. QO, the effect remains on a very low and non-significant level. This result corroborates the results of the power analysis mentioned above. Overall, the three tests clearly demonstrate that the main effects of our manipulations are robust and reliable, and, thus, strongly support our conclusions about the importance of downbeat delays for the swing feel.

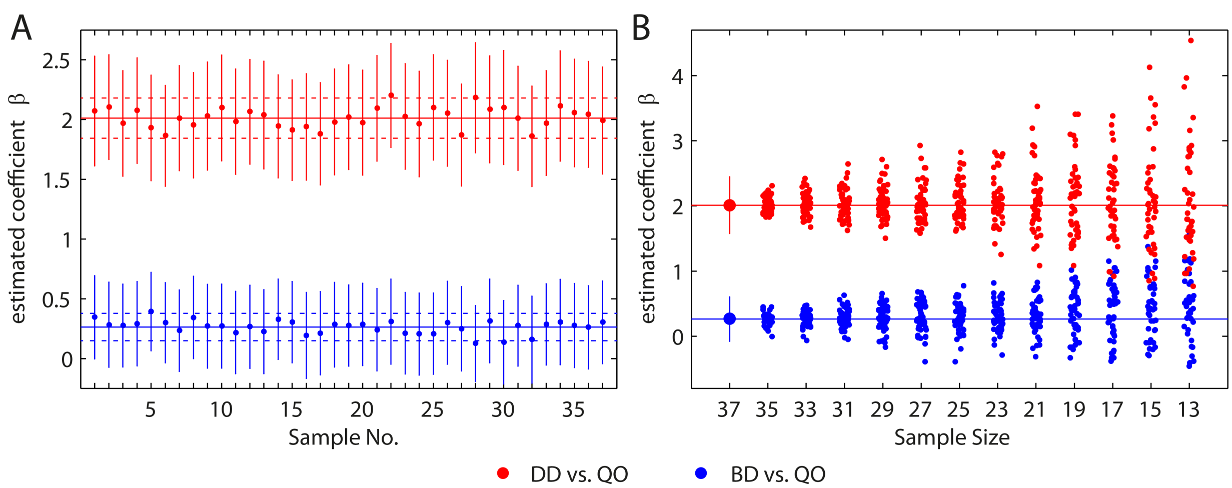


Figure 11: **Effect of outliers on the estimated effects and robustness with respect to sample size.** A) Estimated β coefficients (y-axis) of 37 ordinal regressions (x-axis), in each of which one different participant was excluded for DD vs. QO (red) and BD vs. QO (blue). Solid lines represent the β values for the whole sample as reported in the main analysis. Dashed lines represent +/- 2 standard deviations. B) Results of ordinal regressions using different sample sizes with n=50 iterations each with randomly drawn samples. The horizontal lines show the β coefficient for the whole sample as obtained in the main analysis (N=37) and error bars depict standard errors (SE) of the regression coefficients of the whole sample. These figures clearly demonstrate the robustness of the observed effects.

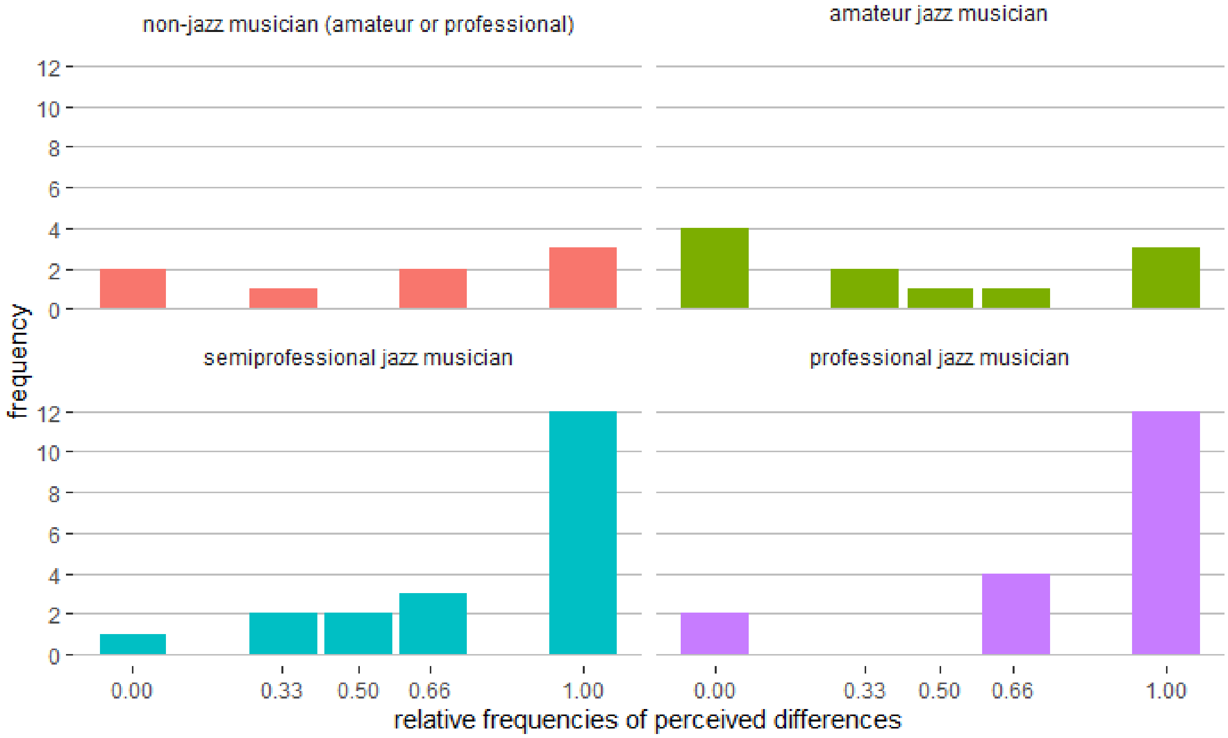


Figure 12: **Perceived differences between different manipulations shown separately for different groups of musicians.** On the horizontal axis relative frequencies are shown, which indicate how often participants perceived differences between manipulations. A relative frequency of 0 means no differences were perceived comparing the different manipulations, 1 means that differences between all manipulations were perceived considering the total number of pieces heard. The vertical axis shows the number of participants for each relative frequency. Visual inspection reveals that non-jazz and amateur jazz musicians had difficulties discriminating between the different manipulations, while both professional and semiprofessional jazz musicians perceived differences very often, i.e. in a majority for every version of every piece.

2.3 Participants' ability to hear differences

In the experiment participants were asked to rate the swing feel of every manipulated version of a piece. Different ratings indicate that participants clearly heard differences between the versions. However, participants may have given the same rating to different versions even when they perceived differences. Therefore participants were always asked whether they perceived any difference, when they gave the same rating to two or more versions of the same piece. The results for the perceived differences are shown in Fig. 12. On the abscissa, the relative frequency of perceived differences is shown. Ideally, differences should always be perceived (i.e., 100%), as the three presented versions differed from each other. Relative frequencies lower than 1.0 indicate that differences between versions were not heard. On the ordinate, the number of participants that perceived the respective percentage of differences is shown. A large number of semiprofessional and professional jazz musicians perceived differences between all versions of every piece (i.e., their relative frequency of perceived differences was 1.0). By contrast, a large fraction of the non-jazz and amateur jazz musicians had difficulties discriminating between the different manipulations. Less than half of them heard differences between all versions. These observations were a reason to restrict our analyses to semiprofessional and professional jazz musicians only.

Condition	β	SE	p	OR	CI
DD vs **QO** (condition 1)	1.229	0.300	$< .001$***	3.42	[1.90 ; 6.16]
BD vs **QO** (condition 2)	0.427	0.295	.148	1.53	[0.86 ; 2.73]
semi-pro vs **pro** (muscat)	0.757	0.530	.153	2.13	[0.76 ; 6.02]
condition 1 × muscat	0.798	0.584	.171	2.22	[0.71 ; 6.97]
condition 2 × muscat	0.456	0.587	.437	1.58	[0.50 ; 4.98]
Threshold coefficients		Estimate	Std. error		
1|2		-1.447	0.333		
2|3		1.194	0.332		
3|4		3.498	0.421		

Table 3: Proportional odds mixed model for groove ratings fitted with Laplace approximation. *** $p < .001$, * $p < .05$. β = regression coefficient. SE = Standard Error. For other abbreviations please refer to Table 1 of the main text.

2.4 Groove ratings

In addition to the swing feel, participants rated the perceived groove for each manipulation and piece in the online experiment. Overall the groove ratings show similar but less clear-cut results than the swing ratings. The average distribution of groove ratings across the pieces shows that the *downbeat delayed* manipulation obtained higher ratings than the *quantized original* and the *both delayed* manipulation. The *quantized original* obtained the lowest ratings (Fig. 13). These effects were confirmed by a logistic ordinal regression, which yielded a significant difference between the *downbeat delayed* and the *quantized original* version but no significant effect between the *both delayed* and *quantized original* versions (Table 3). No difference was found between different jazz musician groups and interactions between musician group and manipulations did not show up. Thus, delaying downbeats and synchronizing offbeats seems to have a positive impact also on the groove feeling of listeners compared to the *quantized original* version. But this impact seems to be smaller for the groove feel than for the swing feel. While the *downbeat delayed* manipulation received 60% of the high swing ratings (3 and 4), the other two manipulations received ratings of 3 or 4 only in 26.4% and 35.4% of the cases, respectively. In contrast, for groove the results for all three manipulations show these high ratings only in the minority of cases (*downbeat delayed*: 41.2%, *quantized original*: 16.5%, *both delayed*: 35.1%).

A similar picture emerges from the ROC-analysis. Fig. 14 displays ROC-curves for the groove ratings for all three pieces. The area under the curve (AUC) reflects the difference in groove ratings between two manipulations. Values of AUC > .5 and a confidence interval (CI) that does not include 0.5, would imply a significantly better groove feel in one manipulation than in the other. For the comparison of *downbeat delayed* vs. *quantized original* this is the case only for "Texas blues" and "Jordu", for "The smudge" the CI includes the value 0.5. In the comparison of *downbeat delayed* vs. *both delayed* only "The smudge" yielded significantly better groove ratings. In contrast, the ROC-analysis for the swing feel consistently yielded a better swing feel in the *downbeat delayed* manipulation compared to the *quantized original* and the *both delayed* manipulation (see Fig. 5 in the main text). Thus, the effect of the manipulations on the groove feel seems to be smaller and/or inconsistent across pieces (in contrast to the effects on the swing feel). Overall these effects might be interpreted as evidence that swing and groove may be dissociable phenomena that are affected by different variables.

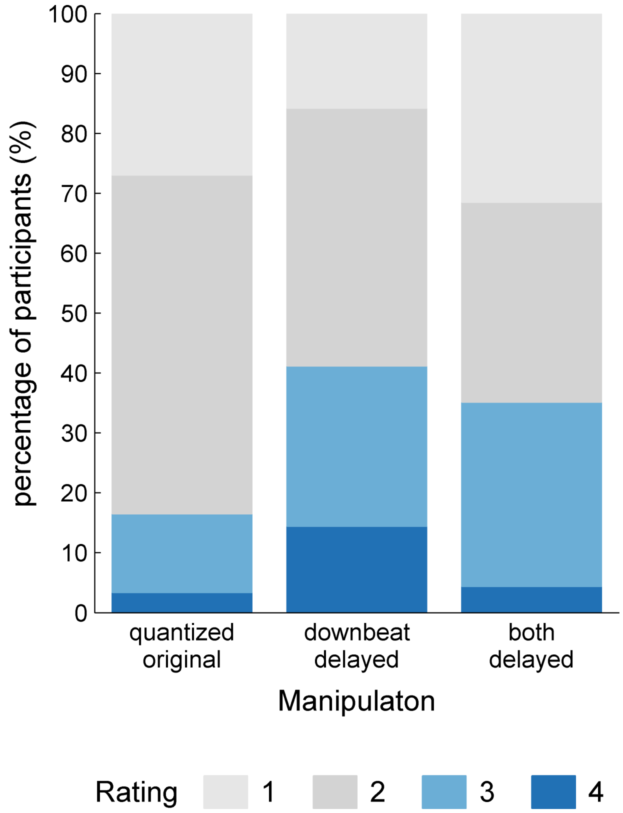


Figure 13: **Distribution of groove ratings given by professional and semi-professional jazz musicians to different manipulated versions.** The three stacked histograms display the proportions of different possible ratings 1-4.

3 Serenade to a cuckoo: second experiment testing different swing ratios

In the framework of this study, we conducted an online experiment (as described in the main text) using manipulated musical extracts to probe the effects of downbeat delays. The first step of the manipulation was to quantize the pieces used in the experiment to a defined *swing ratio* in order to suppress random microtiming deviations that could influence the results. To choose this swing ratio, we oriented ourselves to the results presented in section 1 of the main text (see Fig. 2). Since the swing ratio vs tempo relationship presents large fluctuations, we were not certain about the optimal choice of the swing ratio and initially tried not to deviate too much from the original swing ratio of the recording. For three of the pieces used in the main study ("The smudge", "Texas blues" and "Jordu") the soloist's original swing ratio was not too far from an "optimal" swing ratio (see section 1 of the main text). For "Serenade to a cuckoo", however, the swing ratio of the soloist in the underlying recording was somewhat higher than optimal (1.9) for the tempo of 140 bpm. In the *downbeat delayed* manipulation (which increases swing ratios of the rhythm section) this led to a very high swing ratio of the rhythm section of 2.91 (i.e. nearly punctuated 8th and 16th notes). A swing ratio this high appeared to be detrimental for the swing feel, as revealed by comments from several participants of our experiment. We therefore conducted a separate online experiment in order to compare a version, where the soloist's swing ratio in the *quantized original* is $r_{original} = 1.9$ ("Serenade 1"), with a version, where $r_{original} = 1.6$ ("Serenade 2"), i.e. near the optimum for that tempo according to Fig. 2 of the main text. Details on the two versions of "Serenade to a cuckoo" used in our successive experiments are

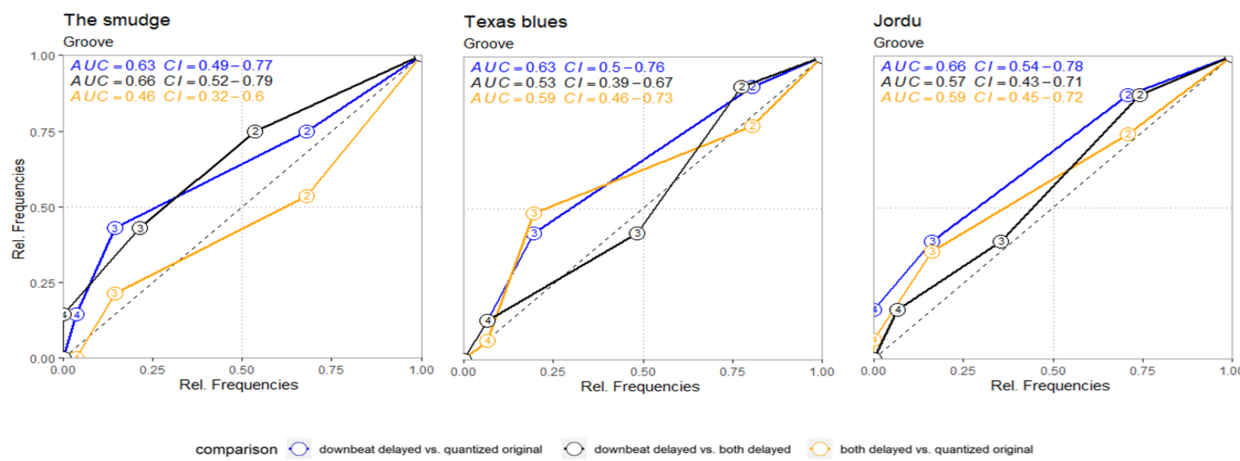


Figure 14: **Receiver Operating Characteristic (ROC) curves for the groove ratings of three pieces.** These curves compare cumulative proportions of the ratings 4 to 1 for two conditions mapped along the horizontal and the vertical axis, i.e. two of the stacked histograms of Fig. 13 are plotted against each other along an axis each. A deviation from the diagonal to either side indicates higher groove ratings for one of the conditions and shows that listeners discriminate between the versions. The area under the curve (AUC) quantifies the deviation from the diagonal ($AUC = 0.5$ means no discrimination). Statistical significance is confirmed, if the AUC confidence intervals (CI) do not contain $AUC = 0.5$.

provided in Table 4. This follow-up study comprised 30 semiprofessional and professional musicians (M: age = 45.40; SD: age = 11.93). Apart from the swing ratio, this online study on "Serenade" remained identical to the one presented in the main text. Fig. 15 shows the distributions of swing ratings for both versions of "Serenade".
With a swing ratio of $r_{original} = 1.9$ the swing ratings do not differ much between the three manipulations. With a swing ratio of $r_{original} = 1.6$, however, the swing ratings for the *downbeat delayed* versions are substantially higher than for both other manipulations. This data pattern is very similar to our findings in the main study.

In order to analyze the effect of swing ratios on the discriminability of our three manipulations, we carried out an ordinal regression with the predictors conditions and versions (i.e. Serenade 1 vs. Serenade 2) and their interaction (Table 5). The first interaction shows that the optimized swing ratio in version Serenade 2 significantly improved the discriminability of the *downbeat delayed* version w.r.t. the *quantized original* version ($OR = 8.00$, $p = 0.004$). Similarly as for the three songs analyzed in the main text, discriminating between the *both delayed* and the *quantized original* manipulations is also impossible here for both versions of Serenade (no main effect of condition 2, $OR = 1.58$, $p = 0.192$, and no interaction, $OR = 1.26$, $p = 0.740$).

The ROC curves for the Serenade 1 version and the Serenade 2 version (Fig. 16) also show a much better discriminability between the manipulations for Serenade 2. A preference for the *downbeat delayed* manipulation is not observable for the ill-chosen swing ratio of Serenade 1, but stands out for the optimized swing ratio of Serenade 2 and is clearly significant as indicated by the large areas under the curve (AUC). For Serenade 2 we find $AUC = 0.74 \pm 0.12$ for the *downbeat delayed* vs. *quantized original* manipulation and $AUC = 0.70 \pm 0.12$ for the *downbeat delayed* vs. *both delayed* manipulation. The corresponding AUC confidence intervals are sufficiently distant from 0.5 and thus underline the significance of the effect. We thus conclude that the

Recording	tempo	$r_{original}$	$r_{soloist}$	r_{rhythm}
Serenade 1	140	1.91	1.9	2.93
Serenade 2	140	1.91	1.6	2.36

Table 4: **Characteristic parameters of two versions of "Serenade to a cuckoo".** The tempo is expressed in beats per minute (bpm). $r_{original}$ stands for the averaged swing ratio of the original piece before manipulation. $r_{soloist}$ represents the swing ratio of the soloist after quantization is applied and r_{rhythm} gives the swing ratio of the rhythm group after the offbeats are synchronized with the soloist in the *downbeat delayed* manipulation.

swing ratio is an essential parameter which must be well-chosen before the effect of different microtiming manipulations on the swing feel can be studied.

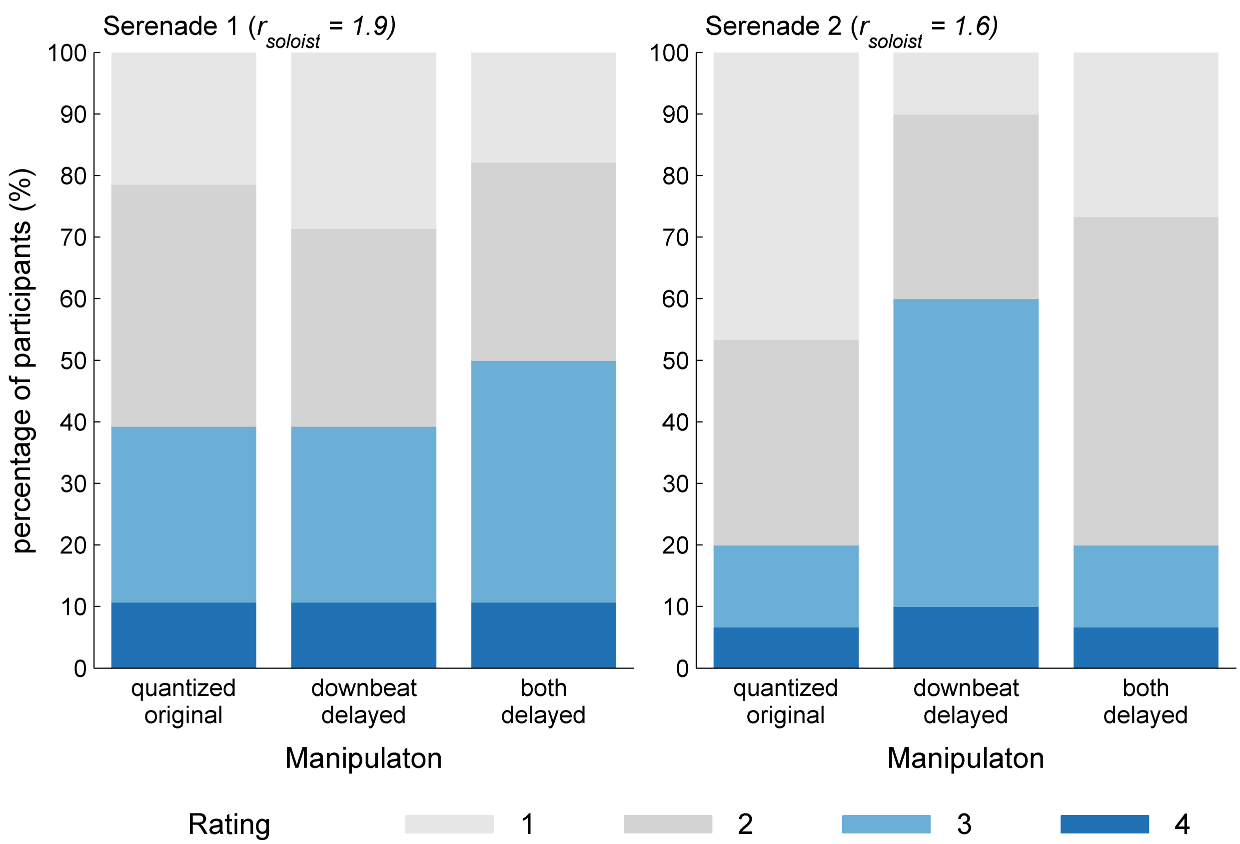


Figure 15: **Distribution of swing ratings for two different versions of "Serenade to a cuckoo".** For "Serenade 1" ($r_{soloist} = 1.9$) the distributions of swing ratings do not differ substantially between the three manipulations. For "Serenade 2" ($r_{soloist} = 1.6$), the *downbeat delayed* manipulation obtained much higher ratings than the *quantized original* and *both delayed* manipulations. The effects for "Serenade 2" are thus nearly identical to our findings for the three pieces in the main text.

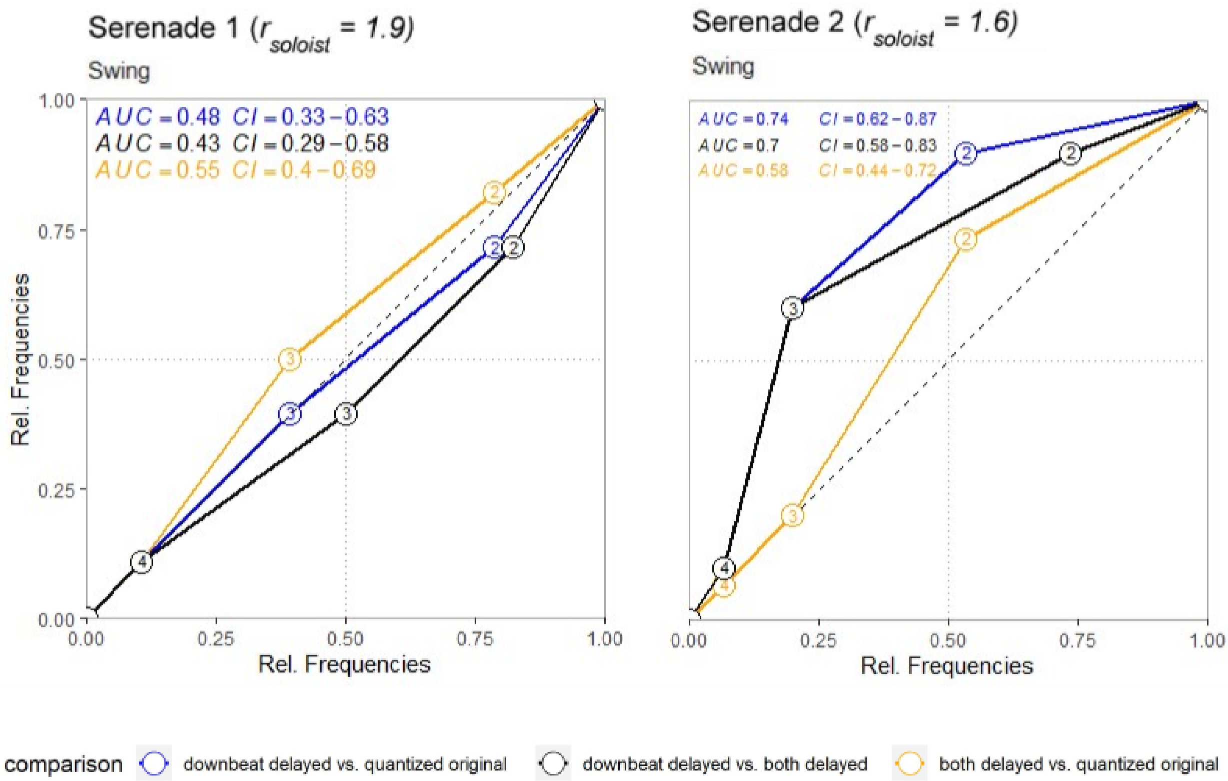


Figure 16: **Receiver Operating Characteristics (ROC) curves for swing ratings of two different versions of "Serenade to a Cuckoo" distinguished by the soloist's swing ratio.** These curves compare cumulative proportions of the ratings 4 to 1 for two conditions mapped along the horizontal and the vertical axis, i.e. two of the stacked histograms of Fig. 15 are plotted against each other along an axis each. A deviation from the diagonal to either side indicates higher swing ratings for one of the conditions and shows that listeners discriminate between the conditions. The area under the curve (AUC) quantifies the deviation from the diagonal ($AUC = 0.5$ means no discrimination). Statistical significance is confirmed by the AUC confidence intervals (CI), which do not contain $AUC = 0.5$. While "Serenade 1" did not produce significant results (AUC confidence interval does include 0.5), for "Serenade 2" the preference for the *downbeat delayed* manipulation is clearly significant (AUC confidence interval does not include 0.5) in comparison with the *quantized original* and the *both delayed* manipulations. The striking difference between "Serenade 1" and "Serenade 2" underlines the importance of choosing an optimized swing ratio before carrying out and comparing different microtiming manipulations.

Condition	β	SE	p	OR	CI
DD vs **QO** (condition 1)	0.919	0.366	.012*	2.51	[1.22 ; 5.13]
BD vs **QO** (condition 2)	0.460	0.352	.192	1.58	[0.79 ; 3.16]
Serenade 1 vs **Serenade 2** (version)	-0.368	0.359	.306	0.69	[0.34 ; 1.40]
condition 1 $\times$ **version**	2.080	0.729	.004**	8.00	[1.92 ; 33.39]
condition 2 $\times$ **version**	0.232	0.701	.740	1.26	[0.31 ; 4.99]

Table 5: Odds mixed model fitted with Laplace approximation comparing two different underlying QO versions of "Serenade to a Cuckoo" as starting points. The two QO versions differed only in the soloist's swing ratio (1.9 for "Serenade 1" and 1.6 for "Serenade 2"). ** $p < 0.01$, * $p < 0.05$. The following abbreviations were used: DD stands for *downbeats delayed*, QO for *quantized original* and BD for *both delayed*. As reference the *quantized original* condition and the "Serenade 2" version were chosen. According to the interactions the discriminability of the different manipulations was significantly higher for "Serenade 2" (swing ratio = 1.6) than for "Serenade 1" (swing ratio = 1.9).

References

[1] Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 42(2):109–127, 1980.

[2] Rune Haubo B Christensen. Cumulative link models for ordinal regression with the r package ordinal. *Journal of Statistical Software*, 2018.

[3] Institute for Digital Research & Education. Ordinal logistic regression— r data analysis examples, 2020.