Improved Listening Experiment Design for Generative Systems

Jeff Ens and Philippe Pasquier [⋆]

Simon Fraser University
jeffe@sfu.ca pasquier@sfu.ca

Abstract. Designing robust listening experiments is a critical component of research on generative music systems, as they are often the primary mechanism by which systems are bench-marked. However, the field lacks a set of guidelines for designing these types of experiments. In order to provide substantiated recommendations for experimental design, we examine the role of two parameters: the proportion of questions and the proportion of participants, both of which are measured relative to the total number of observations. Somewhat surprisingly, these parameters vary significantly from study to study, demonstrating a lack of consensus within the research community. Using experimental data collected from previous studies, we compare the power and reliability of various experimental designs, and arrive at guidelines regarding these proportions.

Keywords: Evaluation, Methodology, Generative Systems

1 Introduction

When evaluating a generative music system (audio or symbolic), human-based assessments are considered the gold standard. In most cases, participants are provided with one or more musical excerpts, and are asked to rate or rank the provided excerpts based on their quality. However, there are no generally accepted guidelines or recommendations for the design of these studies, which is directly evidenced by a high level of variance in experimental designs across studies published in recent years. We use experimental evidence and theoretical reasoning to critically evaluate the design of previously published experiments, and rationalize recommendations for improved experimental design.

We make the distinction between four different methodologies for quantitatively evaluating generative musical systems via a listening test. A modified Turing test (Turing, 2009) can take two forms, one where participants are asked whether a single musical excerpt is computer-generated or human-composed (I) (Hadjeres, Pachet, & Nielsen, 2017; Thickstun, Harchaoui, Foster, & Kakade, 2018; Donahue, Mao, Li, Cottrell, & McAuley, 2019), and another where participants select the human-composed musical excerpt from a pair of musical excerpts (II) (Liang, Gotham, Johnson, & Shotton, 2017). Another approach (III)

[⋆] We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), and the Helmut & Hugo Eppich Family Graduate Scholarship

tasks participants with selecting the higher quality excerpt from a pair of musical excerpts (Huang et al., 2019; Huang, Cooijmans, Roberts, Courville, & Eck, 2017; Hawthorne et al., 2019; Roberts, Engel, Raffel, Hawthorne, & Eck, 2018), where one or more of the excerpts is computer-generated. The final method involves computing the average rating for excerpts from each source (IV) (Collins & Laney, 2017; Collins, Laney, Willis, & Garthwaite, 2016; Pearce & Wiggins, 2007). Note that we use the term source here rather than generative system, as real data is often included as a condition in the experiment.

There are many factors which influence the outcome of a listening experiment. These include, the cultural background of the participant (Eerola, Himberg, Toiviainen, & Louhivuori, 2006), the listening equipment used in the study, and the physical condition of the participant. However, many of these factors are out of the experimenters control. Here we focus on two hyper-parameters which can be directly controlled by the experimenter, the proportion of questions and the proportion of participants. Consider an experiment $\mathcal{E} = \{(Q^{\gamma_1}, S^{\alpha_1}, R^1), ..., (Q^{\gamma_{n_{\texttt{obs}}}}, S^{\alpha_{n_{\texttt{obs}}}}, R^{n_{\texttt{obs}}})\}$ consisting of $n_{\texttt{obs}}$ observations, given a set of questions $Q = \{Q^1, ..., Q^{n_{\texttt{ques}}}\}$ and a set of participants $S = \{S^1, ..., S^{n_{\texttt{par}}}\}$. Note that a question is simply a set of musical excerpts sampled from one or more sources, from which a participant must formulate a response. Given $\mathcal{E}$, the proportion of participants is $n_{\texttt{par}}/n_{\texttt{obs}}$ and the proportion of questions is $n_{\texttt{ques}}/n_{\texttt{obs}}$. Although these hyper-parameters play a significant role, they have not been thoroughly scrutinized in this context.

2 Experimental Design

An experiment is comprised of factors, which are simply independent variables that are manipulated by the experimenter. There are two types of factors: fixed factors, which have a fixed number of levels that are of interest to the researcher; and random factors, where a random subset of the large number of possible levels that are of interest to the researcher are included in the experiment. Typically, in a listening experiment evaluating generative systems, there is one fixed factor, where each level is a different source (i.e. a generative system or real data). The participant factor, is a well-known random factor included in most experiments. Practical limitations place restrictions on the total number of levels (i.e. participants) that can be feasible included in the experiment, which forces the experimenter to randomly sample from the participant population of interest. In experiments that evaluate generative systems, there is another important random factor, the questions. It is clearly impossible to include all possible questions within an experiment, so we must settle for a random sample of questions.

Once we have established the factors within an experiment, it is necessary to determine the experiment design, which specifies the relationship between factors. Pairs of factors can be crossed or nested. If two factors are crossed, every level of one factor co-occurs with every level the other factor. If factors are nested, each level of a factor co-occurs with only one level of the other. For example, consider a methodology I experiment conducted with two participants (S^1, S^2), comparing two sources (M^1, M^2), where 2 excerpts $(e_1^{M^k}, e_2^{M^k})$ are generated from each source M^k. Since methodology I asks participants to listen to a single excerpt and predict whether it was computer-generated or human-composed, we have two unique questions ($Q_1^{M^k}, Q_2^{M^k}$) per source, where each question consists of a single musical excerpt. Here, the question factor is nested within the source factor, as each question $Q_i^{M^k}$ is unique to the source M^k. Note that this is the case for all listening experiments evaluating generative systems, since it is exceedingly rare to sample the same question from two different sources. If each of the 4 questions are shown to each participant, then the participant factor would be crossed with the question factor, as each participant-excerpt combination $(S^i, e_j^{M^k})$ is part of the experiment.

There are three common experimental designs: crossed-question, partially-crossed-question and nested-question. A crossed-question design, shown in Figure 1a, exposes each participant to the same set of questions. A nested-question design, shown in Figure 1c, nests questions within participants, so that each participant is exposed to a different set of questions. It is also possible to employ a partially-crossed-question design, shown in Figure 1b, where the set of questions that each participant is exposed to is randomly drawn from a set of questions. As a result, participants will sometimes be exposed to the same question. Although 6 observations are collected in each of the experimental designs shown in Figure 1, the sample size of the question random factor varies. Consequently, the proportion of questions is smallest for a crossed-question design ($\frac{n_{\text{ques}}}{n_{\text{obs}}} = \frac{2}{6}$) and largest for a nested-question design ($\frac{n_{\text{ques}}}{n_{\text{obs}}} = \frac{6}{6}$).

	$Q_1^{M_1}$	$Q_2^{M_2}$
S^1	x	x
S^2	x	x
S^3	x	x

(a)

	$Q_1^{M_1}$	$Q_2^{M_1}$	$Q_3^{M_2}$	$Q_4^{M_2}$
S^1	x	-	x	-
S^2	-	x	x	-
S^3	-	x	-	x

(b)

	$Q_1^{M_1}$	$Q_2^{M_1}$	$Q_3^{M_1}$	$Q_4^{M_2}$	$Q_5^{M_2}$	$Q_6^{M_2}$
S^1	x	-	-	x	-	-
S^2	-	x	-	-	x	-
S^3	-	-	x	-	-	x

(c)

Fig. 1: Three different experimental designs: crossed-question (a), partially-crossed-question (b) and nested-question (c). The cells with x denote the observations that are collected.

For purposes of conceptual clarity, we only consider the case where an experiment consists of ≤ 2 sources, as this is an atomic unit that larger experiments are easily factored into. For example, consider the paired listening experiment presented in the Music Transformer paper (Huang et al., 2019), which compares four sources (Music Transformer, Transformer, LSTM, and the Maestro dataset (Hawthorne et al., 2019)) using methodology III. This can be factored into $6 = \binom{4}{2}$ distinct sub-experiments corresponding to each possible pair of sources. Clearly, if we take steps to improve each sub-experiment, it will have a positive effect on the experiment as a whole.

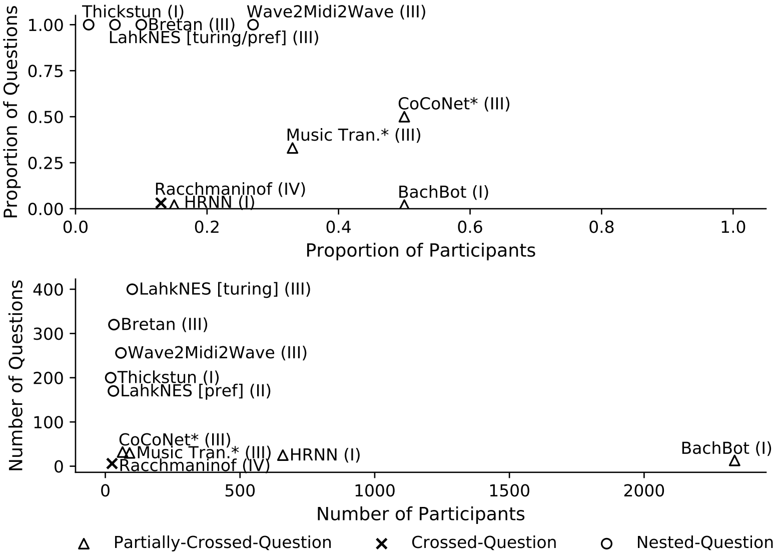

Fig. 2: The experimental designs employed in recent listening studies for generative systems. Stars indicate that the number/proportion of participants could not be calculated exactly.

3 Motivation

There are several motivating factors for this research. First and foremost, without robust experimental design, any claims based on the experimental results are weakened, and in the extreme case completely invalid. Secondly, there are currently no standard recommendations for experimental design, which results significant discrepancies between studies. In Figure 2, we plot the proportion of questions ($\frac{n_{\text{ques}}}{n_{\text{obs}}}$), the proportion of participants ($\frac{n_{\text{par}}}{n_{\text{obs}}}$), the experimental design, and the methodology for several recent listening experiments for which the relevant information was available. Of particular concern, is the fact that the proportion of questions, and experiment design vary significantly across experiments, indicating a lack of consensus amongst the research community. Finally, given the high costs of conducting a study, it is essential that the studies produce accurate results and are implemented to make efficient use of the allocated resources.

4 Experiment 1 : Calculating Experimental Power

A typical approach to evaluate an experimental design is to calculate the power, which is simply the inverse of the probability of Type II error. In order to calculate the power of an experiment, there are two factors which must be considered: the variance components, and the sample size (Judd, Westfall, & Kenny, 2017).

Note that in our case, there is not a single sample size, but rather a sample size for the participant random factor, and a sample size for the question random factor. To explore the differences between nested-question and crossed-question experiment designs, we conduct a parameter sweep for the number of participants, and the number of questions per participants, calculating the power for each pair of parameters. We use power calculations designed for experiments with two random factors (Westfall, Kenny, & Judd, 2014), and compute the variance components from a previous experiment (Collins & Laney, 2017), which featured a crossed-question design. Since power calculations are not available for partially-crossed designs we can not explicitly explore this experiment design here. We deliberately set the x axis of Figure 3 to be the number of questions per participant, rather than the total number of questions, so that the power at each (x,y) coordinate can be directly compared, as the total number of observations is equivalent for each experimental design.

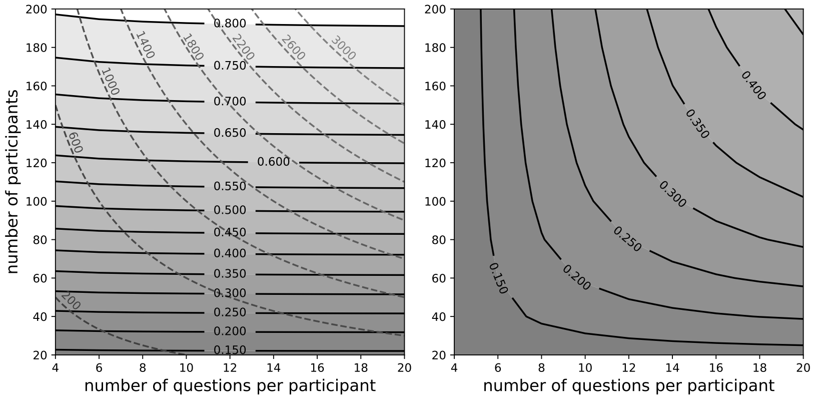

Fig. 3: Power simulation for nested-question (left) and crossed-question (right) experimental designs using variance components estimated from Collins' study. Each dashed line in left plot illustrates the possible combinations of $n_{\texttt{par}}$ and questions per participant given a constant number of observations ($n_{\texttt{obs}}$).

The results in Figure 3 demonstrate that nested designs are uniformly more powerful than crossed designs, as we get an average of 2.3 times more power when using a nested experiment with the same number of total observations. The reason for this is rather straightforward, as a nested-question experiment design can make use of $n_{\texttt{obs}}$ unique questions, while in a crossed design we are restricted to $\frac{n_{\texttt{obs}}}{n_{\texttt{par}}}$ unique questions. Provided that variance components related to the question factor are non-zero and $n_{\texttt{par}} \geq 1$, nested-question experiments will always be more powerful than crossed question experiments, as they increase the sample size for the question factor by a factor of $n_{\texttt{par}}$ (Judd et al., 2017).

The dashed lines in Figure 3a show the different possible combinations of participants ($n_{\texttt{par}}$) and questions per participant given a constant number of observations ($n_{\mathtt{obs}}$). This reveals that the power decreases when we decrease the proportion of participants $\frac{n_{\mathtt{par}}}{n_{\mathtt{obs}}}$, while holding the proportion of questions constant ($\frac{n_{\mathtt{ques}}}{n_{\mathtt{obs}}} = 1$). Note that we cannot observe this same effect in Figure 3b, since changes to the proportion of participants are confounded with changes to the proportion of questions. The same type of effect can be observed in a nested-participant experiment design, where the power decreases when the proportion of questions decreases. However, this type of experiment design is highly impractical as it requires collecting a single response from each participant.

We can also observe that in a crossed design, there is little advantage to increasing the number of participants or increasing the number of questions per participant separately. Power mainly increases when the number of participants and the number of questions per participant are increased together. Furthermore, it is possible to reach a point when adding an additional participant has no effect on the power at all, since the contour lines eventually become almost vertical. In contrast, when using a nested design most of the gains come from adding participants, since this effectively increases the total number of questions in the experiment, as the questions at each participant level are unique. The efficiency of the nested-question experiment design, is that it allows for both the sample size for participants and questions to be increased simultaneously. Collectively, these results demonstrate that crossed-question experiments are under-powered, and that decreasing the proportion of participants or the proportion of questions decreases the power.

5 Experiment 2 : Simulating Inter-Experiment Variance

In this experiment, we aim to measure inter-experiment variability, quantifying the reliability of different experimental designs. Formally, given an experiment $\mathcal{E}$, let $\psi^k_{n_{\mathtt{par}},n_{\mathtt{ques}}}$ denote the result (i.e. the average score for a source) for a randomly sampled subset of $\mathcal{E}$, containing k observations, $n_{\mathtt{par}}$ participants, and $n_{\mathtt{ques}}$ questions, where each each observation in $\psi^k_{n_{\mathtt{par}},n_{\mathtt{ques}}}$ involves the same source(s). To observe the difference between two experimental designs (α and β), we compute $\psi^k_{n^\alpha_{\mathtt{par}},n^\alpha_{\mathtt{ques}}}$ and $\psi^k_{n^\beta_{\mathtt{par}},n^\beta_{\mathtt{ques}}}$ r times, resulting in the sets Ψ^α and Ψ^β. Then we use Levene's test (Levene, 1960) to determine if the variance of Ψ^α and Ψ^β differs significantly. The entire procedure is repeated 100 times with $r = 50000$, producing 100 p-values. In order to be sure that our results are simply not an artifact of the sub-experiment sampling procedure, we also conduct the same procedure using a version of the data where the responses have been randomly sampled from a uniform distribution, counting the proportion of times that $\sigma(\Psi^\alpha) < \sigma(\Psi^\alpha)$, where $\sigma(\Psi^i)$ denotes the variance of the set Ψ^i. If the sub-experiment sampling procedure has a significant effect on the outcome, we would expect this proportion to vary significantly from 0.5, a hypothesis which can be tested using the Binomial test.

We use the experimental results provided by the authors of following four listening experiments: BachBot (Liang et al., 2017), Wave2Midi2Wave (Hawthorne et al., 2019), LahkNES (Donahue et al., 2019), and Racchmaninoff (Collins & Laney, 2017). Although we contacted the authors of 15 different studies, we only received experimental results from the four listed above. Note that the experimental design of the original experiments will place inherent limitations on the types of simulations that we can conduct. In the BachBot study, each participant is presented with two different questions, randomly selected from a pool of 13 questions, resulting in a partially-crossed-question design. With this data, we can simulate a partially-crossed-question design $\psi^{10}_{10,2}$ and a nested-question design $\psi^{10}_{10,10}$. In the Wave2Midi2Wave and LahkNES studies, there are almost no duplicate questions in the entire experiment, which only allows us to manipulate the proportion of participants. We simulate two nested-question designs: $\psi^{10}_{5,10}$, and $\psi^{10}_{10,10}$. Using the Racchmaninoff data, we can simulate a crossed-question design $\psi^{9}_{3,3}$ and a partially-crossed-question design $\psi^{9}_{9,3}$. For each comparison, the proportion of significant results after applying the false discovery rate correction (Benjamini & Yekutieli, 2001) is shown in Table 1. The Binomial test for the Wave2Midi2Wave simulations was significant, indicating that the sub-experiment sampling procedure biased the result, so this simulation was excluded from the results. In all other cases, the Binomial test was insignificant. Collectively, the results demonstrate that increasing the proportion of participants or questions decreases the inter-experiment variance, confirming the theoretical results presented in experiment 1.

data source	k	n^{α}_{par}	n^{α}_{ques}	n^{β}_{par}	n^{β}_{ques}	proportion of significant trials
LahkNES [pref]	10	5	10	10	10	1.00
LahkNES [turing]	10	5	10	10	10	1.00
BachBot	10	10	2	10	10	1.00
Racchmaninoff	9	3	3	9	3	.95

Table 1: The proportion of trials where Ψ^{α} exhibits more variance than Ψ^{β}.

6 Discussion and Recommendations

In addition to considering the power and reliability of a particular experimental design, it is also worth taking the end-point of the experiment into account. In most cases, the end-point of a listening study for the evaluation of generative systems is an average score for each system. In contrast, the endpoint of an experiment measuring the valence and arousal of audio clips, is the average valence and arousal for each audio clip. There is a subtle difference between these two types of experiments. In the first experiment, audio excerpts are a random factor, where we take a random sample from the entire population of possible generated excerpts. In the second experiment, audio excerpts are a fixed factor, where we are interested only in the levels contained within the experiment. We do not expect that the results for one particular audio excerpt will generalize to another audio excerpt in the second experiment. As a result, it makes sense to collect multiple observations from multiple participants for each audio excerpt, as we need the average response for each excerpt to be reflective of how the entire participant population feels about that excerpt. However, when conducting a prototypical listening experiment for generative systems, we care about what the entire population thinks of each source, not the individual audio excerpts. To make matters worse, our experiments demonstrated that collecting multiple observations for a single audio excerpt actually makes the results we actually care about less reliable and the experiment as a whole less powerful, as the size of the random sample for audio excerpts representing each source is unnecessarily reduced.

This is not to say that collecting multiple observations for a single audio excerpt is always wasteful. In fact, a crossed-question experiment was necessary for calculating the variance components used our simulations. Furthermore, in cases where inter-rater agreement is the endpoint of an experiment, it is necessary for $\frac{n_{\text{ques}}}{n_{\text{obs}}} < 1$. However, most listening experiments for generative systems do not measure inter-rater agreement. Ultimately, it is absolutely essential that the experiment design matches the goals of the research question, otherwise we often end up needlessly sacrificing power and reliability in our experiments. For those who are conducting a prototypical listening experiment for generative systems, we offer the following advice. Since resources (i.e. time and money) are finite, we will assume that a fixed number of observations ($n_{\texttt{obs}}$) can be collected, irrespective of the experiment design. As our experimental results demonstrate that the sample size of the question and participant random factors have a significant effect on the power and reliability of the experiment, an ideal experimental design will maximize $\frac{n_{\text{ques}}}{n_{\text{obs}}}$ and $\frac{n_{\text{par}}}{n_{\text{obs}}}$. First and foremost, this means it is essential to avoid crossed-question experimental designs, as they reduce $\frac{n_{\text{ques}}}{n_{\text{obs}}}$ by a factor of $n_{\texttt{par}}$. In most cases, there are relatively few barriers to selecting a nested-question design ($\frac{n_{\text{ques}}}{n_{\text{obs}}} = 1$) or a partially-crossed-question design with a large proportion of questions, as sampling from most models is cheap. In fact, we have seen this experimental design employed in several listening studies (Donahue et al., 2019; Thickstun et al., 2018; Hawthorne et al., 2019). However, there are many listening studies which feature a small proportion of questions, needlessly sacrificing power and reliability. Although our results demonstrate that collecting each response from a unique participant ($\frac{n_{\text{par}}}{n_{\text{obs}}} = 1$) would be optimal, this may not be practical, as there are costs associated with obtaining each participant. Fortunately, most experiments do a good job balancing the proportion of participants, collecting a modest amount of responses from each participant.

7 Conclusion

We have examined two critical parameters for the experimental design of listening studies: the proportion of questions $\frac{n_{\text{ques}}}{n_{\text{obs}}}$, and the proportion of participants $\frac{n_{\text{par}}}{n_{\text{obs}}}$. Through experimentation we demonstrated that when $\frac{n_{\text{ques}}}{n_{\text{obs}}} < 1$ or $\frac{n_{\text{par}}}{n_{\text{obs}}} < 1$, the power and reliability of the experiment are reduced. Since listening studies are a fundamental aspect of research involving generative systems, and a consensus on best practices for listening experiment design has yet to emerge, these recommendations will undoubtedly be a useful reference point for future research.

References

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, 1165–1188.

Bretan, M., Weinberg, G., & Heck, L. (2017). A unit selection methodology for music generation using deep neural networks. *Proc. of the International Conference on Computational Creativity.*

Collins, T., & Laney, R. (2017). Computer-generated stylistic compositions with long-term repetitive and phrasal structure. *Journal of Creative Music Systems*, *1*(2).

Collins, T., Laney, R., Willis, A., & Garthwaite, P. H. (2016). Developing and evaluating computational models of musical style. *AI EDAM*, *30*(1), 16–43.

Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W., & McAuley, J. (2019). Lakhnes: Improving multi-instrumental music generation with cross-domain pre-training. In *Proc. of the 20th international society for music information retrieval conference* (pp. 685–692).

Eerola, T., Himberg, T., Toiviainen, P., & Louhivuori, J. (2006). Perceived complexity of western and african folk melodies by western and african listeners. *Psychology of Music*, *34*(3), 337–371.

Hadjeres, G., Pachet, F., & Nielsen, F. (2017). Deepbach: a steerable model for bach chorales generation. In *Proc. of the 34th international conference on machine learning* (pp. 1362–1371).

Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.-Z. A., Dieleman, S., ... Eck, D. (2019). Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proc. of the 7th international conference on learning representations.*

Huang, C. A., Cooijmans, T., Roberts, A., Courville, A. C., & Eck, D. (2017). Counterpoint by convolution. In *Proceedings of the 18th international society for music information conference* (pp. 211–218).

Huang, C. A., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., ... Eck, D. (2019). Music transformer: Generating music with long-term structure. In *7th international conference on learning representations.*

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625.

Levene, H. (1960). Contributions to probability and statistics. *Essays in honor of Harold Hotelling*, 278–292.

Liang, F. T., Gotham, M., Johnson, M., & Shotton, J. (2017). Automatic stylistic composition of bach chorales with deep lstm. In *Proc. of the 18th international society for music information retrieval conference* (pp. 449–456).

Pati, A., Lerch, A., & Hadjeres, G. (2019). Learning to traverse latent spaces for musical score inpainting. In *Proc. of the 20th international society for music information retrieval conference* (pp. 343–351).

Pearce, M. T., & Wiggins, G. A. (2007). Evaluating cognitive models of musical composition. In *Proc. of the 4th international joint workshop on computational creativity* (pp. 73–80).

Roberts, A., Engel, J. H., Raffel, C., Hawthorne, C., & Eck, D. (2018). A hierarchical latent vector model for learning long-term structure in music. In *Proceedings of the 35th international conference on machine learning* (pp. 4361–4370).

Thickstun, J., Harchaoui, Z., Foster, D. P., & Kakade, S. M. (2018). Coupled recurrent models for polyphonic music composition. In *Proc. of the 20th international society for music information retreival conference* (pp. 311–318).

Turing, A. M. (2009). Computing machinery and intelligence. In *Parsing the turing test* (pp. 23–65). Springer.

Westfall, J., Kenny, D. A., & Judd, C. M. (2014). Statistical power and optimal design in experiments in which samples of participants respond to samples of stimuli. *Journal of Experimental Psychology: General*, *143*(5).

Wu, J., Hu, C., Wang, Y., Hu, X., & Zhu, J. (2019). A hierarchical recurrent neural network for symbolic melody generation. *IEEE Transactions on Cybernetics*, *50*(6), 2749–2757.