# Chapter 1

# Large-scale subjective evaluation

[6] addresses difficulty in quantitative evaluation, suggesting the use of a learned critic in a manner similar to GANs [4]. In a later report, [5] attribute difficulty in evaluation due to lack aim: algorithmic composition, design of compositional tools, and computational modelling of musical styles or music cognition all have different motivations and should thus be evaluated differently.

Following advice of [5], we identify our key motivation as algorithmic composition: generation of novel compositions. To evaluate our success, we employ a subjective evaluation method.

[2] criticizes a musical Turing test as providing little data about how to improve the system, suggesting that listener studies using music experts may be more insightful.

## 1.1 Evaluation framework design

### 1.1.1 Software architecture

The frontend utilizes React and Redux, allowing us to collect fine-grained user action data. Azure App Service is used to host an Express web-service which randomizes experimental questions and collects responses. The data is stored to Azure Data Storage and processed in batch MapReduce using Azure HDInsight.

### 1.1.2 User interface

The landing page for http://bachbot.com/ is shown in fig. 1.1.

Clicking "Test Yourself" redirects the participant to a user information form (fig. 1.2) where users self-report their age group prior music experience into the categories shown.


## Challenge description

We will present you with some short samples of music which are either extracted from Bach's own work or generated by BachBot. Your task is to listen to both and identify the Bach originals.

To ensure fair comparison, all scores are transposed to C-major or A-minor and set to 120 BPM.

Fig. 1.1 The first page seen by a visitor of http://bachbot.com

After submitting the background form, users were redirected to the question response page shown in fig. 1.3. This page contains two audio samples, one extracted from Bach and one generated by BachBot, and users were asked to select the sample which sounds most similar to Bach. Users were asked to provide five consecutive answers and then the overall percentage correct was reported.

### 1.1.3 Question generation

Questions were generated for both harmonizations (using the same abbreviations as defined in

fliang: ref

) as well as original compositions (denoted SATB as all parts are generated). For each question, a random chorale was selected without replacement from the corpus and paired with a corresponding harmonization. SATB sampls were paired with chorales randomly sampled

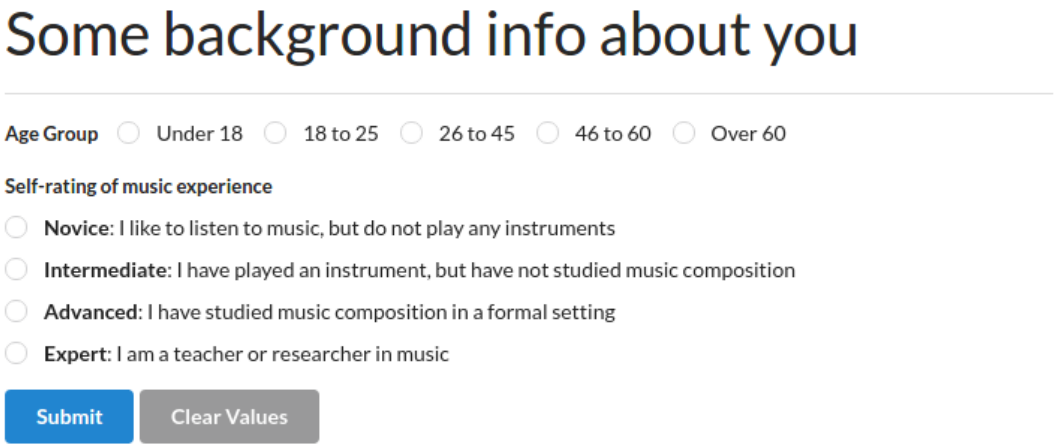



Fig. 1.2 User information form presented after clicking "Test Yourself"

| Question type | # questions available | Expected # responses per participant |
|---|---|---|
| S | 2 | 0.25 |
| A | 2 | 0.25 |
| T | 2 | 0.25 |
| B | 2 | 0.25 |
| AT | 8 | 1 |
| ATB | 8 | 1 |
| SATB | 12 | 2 |

Table 1.1 Composition of questions on http://bachbot.com

from the corpus. The five question answered by any given participant were comprised of one S/A/T/B question chosen at random, one AT question, one ATB question, and two original compositions. See table 1.1 for details.

## 1.2 Results

### 1.2.1 Participant backgrounds and demographics

We recieved a total of

fliang: FILL THIS IN LAST

responses from

fliang: FILL THIS IN LAST

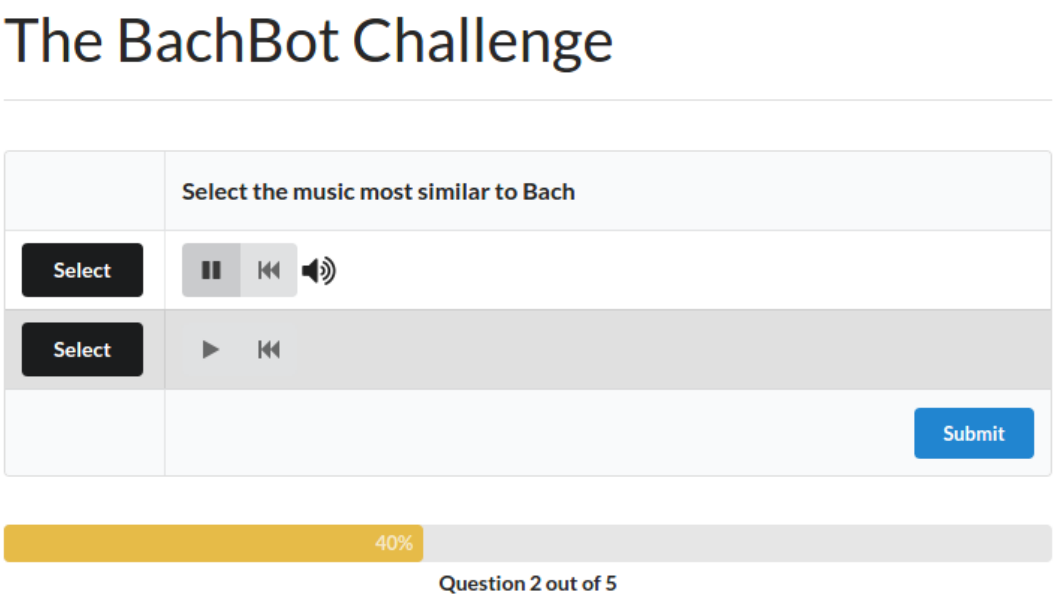



Fig. 1.3 Question response interface used for all questions

different countries. As evidenced by fig. 1.4, our participant is diverse and includes participants from six different continents. fig. 1.5 shows that while the majority of our participants are between 18 – 45 and have played an instrument, more than 20%

fliang: FIX NUMBER LAST

have either formally studied or taught music theory.

### 1.2.2 BachBot's performance results

fliang: fig. 1.6 suggests performance is weakest on harmonizations. Unsurprising because we only do 1-best and don't account for future. Bidirectional LSTM or N-best lattice search (reference marcin) would do better

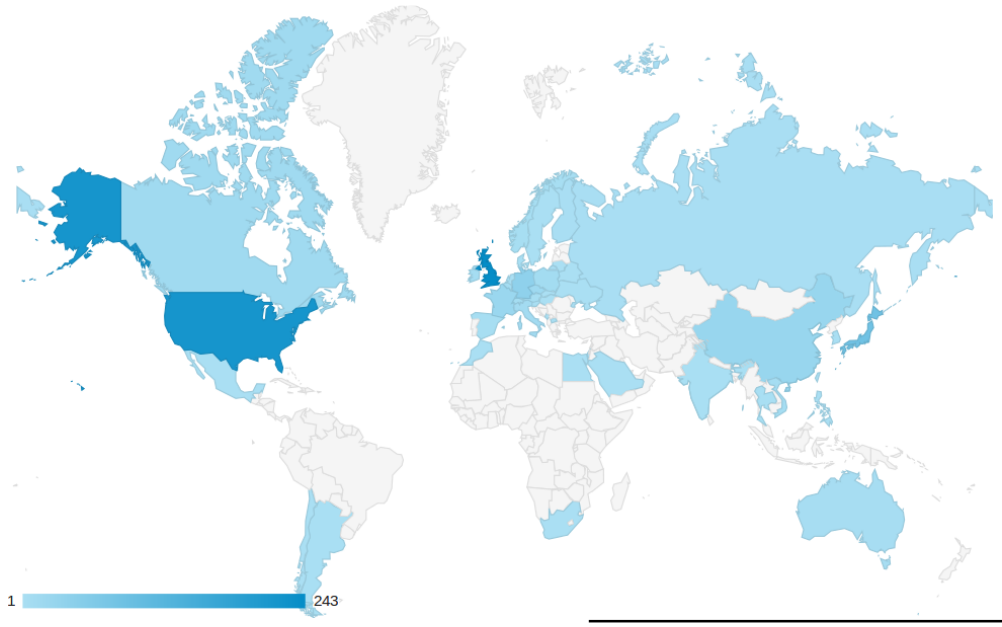
fig. 1.6 shows the performance of BachBot on various question types. It shows that 59%

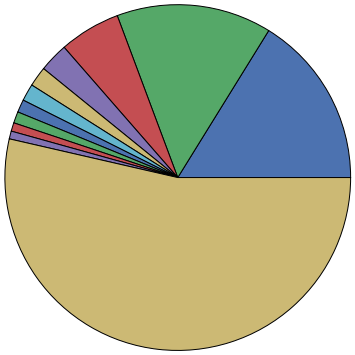
fliang: VERIFY LAST

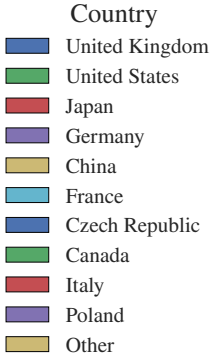
of participants could correctly identify original Bach from BachBot's generated music. As the baseline method of randomly guessing between the two choices in fig. 1.3 achieves 50%, our findings suggest that **the average participant has only a** 9%

**fliang: VERIFY LAST**

**better chance than randomly guessing when distinguishing Bach from BachBot correctly**.

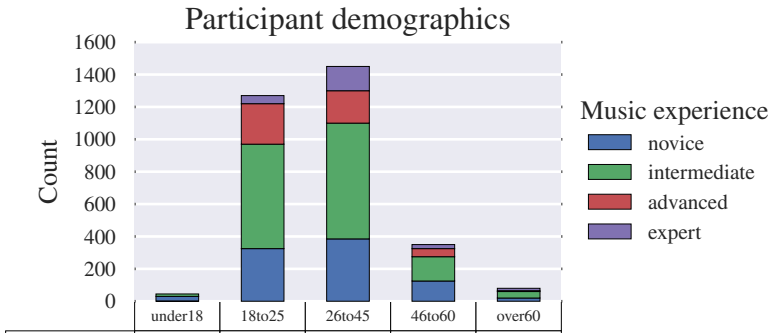




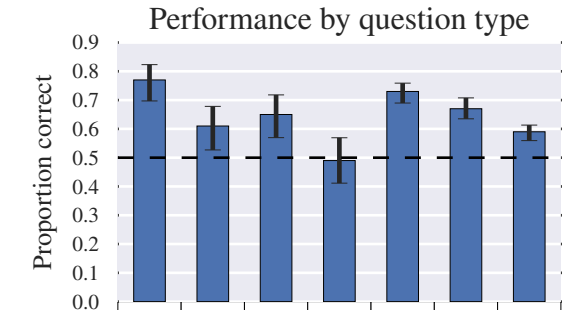

| Country | Responses | Proportion |
|---|---|---|
| UK | 243 | 16.0% |
| US | 218 | 15.0% |
| Japan | 86 | 6.0% |
| Germany | 41 | 3.0% |
| China | 28 | 2.0% |
| France | 24 | 2.0% |
| Czechia | 18 | 1.0% |
| Canada | 16 | 1.0% |
| Italy | 12 | 1.0% |
| Poland | 11 | 1.0% |
| Other | 805 | 54.0% |

Fig. 1.4 Geographic distribution of participants




| | under18 | 18to25 | 26to45 | 46to60 | over60 |
|---|---|---|---|---|---|
| novice | 30.0 | 325.0 | 385.0 | 125.0 | 20.0 |
| intermediate | 15.0 | 645.0 | 715.0 | 150.0 | 40.0 |
| advanced | 0.0 | 250.0 | 200.0 | 50.0 | 5.0 |
| expert | 0.0 | 50.0 | 150.0 | 25.0 | 15.0 |

Fig. 1.5 Demographics of participants




| | S | A | T | B | AT | ATB | SATB |
|---|---|---|---|---|---|---|---|
| Proportion | 0.77 | 0.61 | 0.65 | 0.49 | 0.73 | 0.67 | 0.59 |

Fig. 1.6 responses-Mask

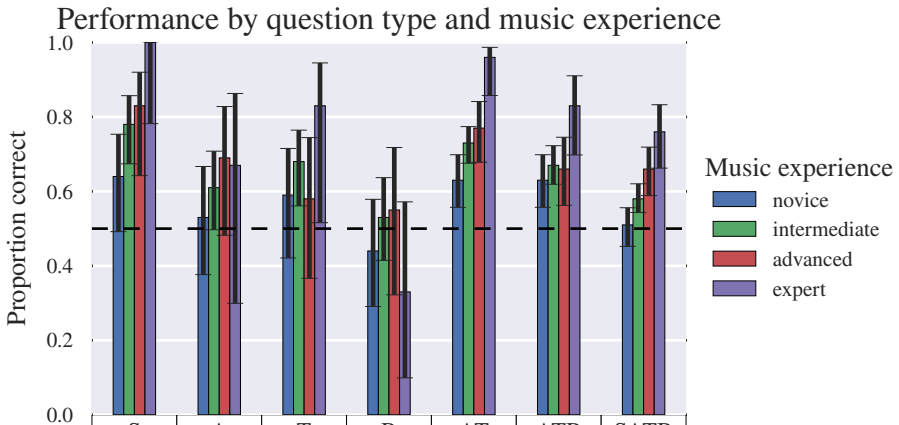
fig. 1.6 also shows that participants had more trouble discriminating entire compositions (SATB) than harmonizations (AT, ATB) where a subset of the parts have already been given. While this may seem counterintuitive, recall that the model in

fliang: reference

is uni-directional and does not account for any future constraints on other parts. We made this design decision intentionally because one of our requirements was sampling the model for novel compositions. However, since harmonization tasks provide the full past and future context for other parts, they effectively impose constraints on LSTM hidden state dynamics. We expect methods which account for both future and past context (e.g. using the output sequence from a bidirectional RNN

fliang: cite

inputs) to mitigate this problem, which we leave for future work.




| | S | A | T | B | AT | ATB | SATB |
|---|---|---|---|---|---|---|---|
| novice | 0.64 | 0.53 | 0.59 | 0.44 | 0.63 | 0.63 | 0.51 |
| intermediate | 0.78 | 0.61 | 0.68 | 0.53 | 0.73 | 0.67 | 0.58 |
| advanced | 0.83 | 0.69 | 0.58 | 0.55 | 0.77 | 0.66 | 0.66 |
| expert | 1.0 | 0.67 | 0.83 | 0.33 | 0.96 | 0.83 | 0.76 |

Fig. 1.7 responses-mask-MusicExperience

When only a only single part is composed by BachBot, we find the results vary significantly across different parts. Composing the soprano part proved to be the easiest to discriminate, an unsurprising result given that in chorale style music soprano parts are responsible for the melody

fliang: cite

. Composing the alto and tenor parts achieved similar performance as composing all four parts, a result which may also be caused by not accounting for future constraints on model outputs. Removing the bass proved to be the most perceptually difficult to discern from real Bach.

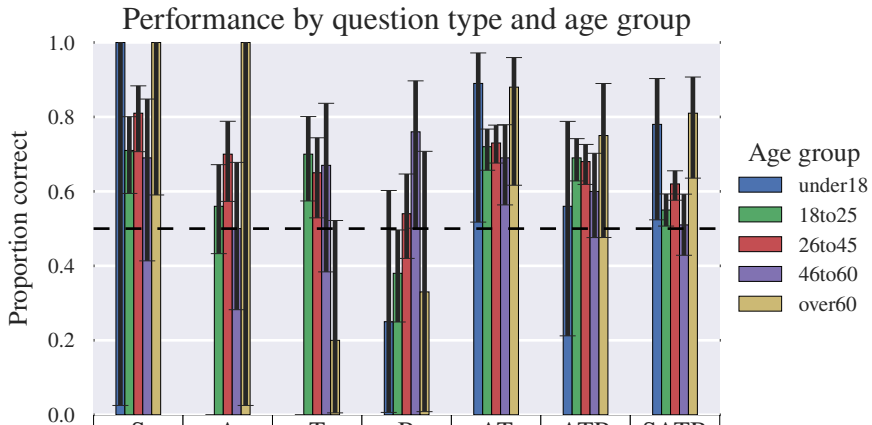
In fig. 1.7, responses are further segmented by music experience. Unsurprisingly, we find that the proportion of correct responses correlates positively with experience.

fig. 1.9 shows the proportion correct for each question. Encouragingly, it shows that 41.7%

fliang: VERIFY LAST

of the SATB pairs were not statistically different than baseline, suggesting that **while not always consistent BachBot is capable of composing music which the average participant cannot discern from actual Bach**.

fliang: Have Mark analyze bad examples in fig. 1.9




| | S | A | T | B | AT | ATB | SATB |
|---|---|---|---|---|---|---|---|
| under18 | 1.0 | 0.0 | 0.0 | 0.25 | 0.89 | 0.56 | 0.78 |
| 18to25 | 0.71 | 0.56 | 0.7 | 0.38 | 0.72 | 0.69 | 0.55 |
| 26to45 | 0.81 | 0.7 | 0.65 | 0.54 | 0.73 | 0.68 | 0.62 |
| 46to60 | 0.69 | 0.5 | 0.67 | 0.76 | 0.69 | 0.6 | 0.51 |
| over60 | 1.0 | 1.0 | 0.2 | 0.33 | 0.88 | 0.75 | 0.81 |

Fig. 1.8 responses-mask-Agegroup

## 1.3 User feedback

The modulations and part writing were the giveaway for me (and once or twice the phrasing)

Got 5/5. The trick is to listen for the unnatural pauses at regular intervals.

Cool project, I scored 100% so I'm quite pleased with myself ;o) I do play an instrument although I'm not classical trained. If I had an inkling to why I could choose the background phrasing of the Bach pieces is far more elegant than the computer generated pieces.

@samim @feynmanliang really impressive! If I didn't know about counterpoint that quiz would've stumped me

## 1.4 Competitive analysis of large-scale evaluation methodologies

fliang: Breakdown costs of Azure CDN, App Service, BlobStore. Most expensive was domain registration

fliang: Compare costs and quality with MTurk

Higher quality. Music experts are not usually doing tasks on MTurk, but would be very interested in an open-source research project.

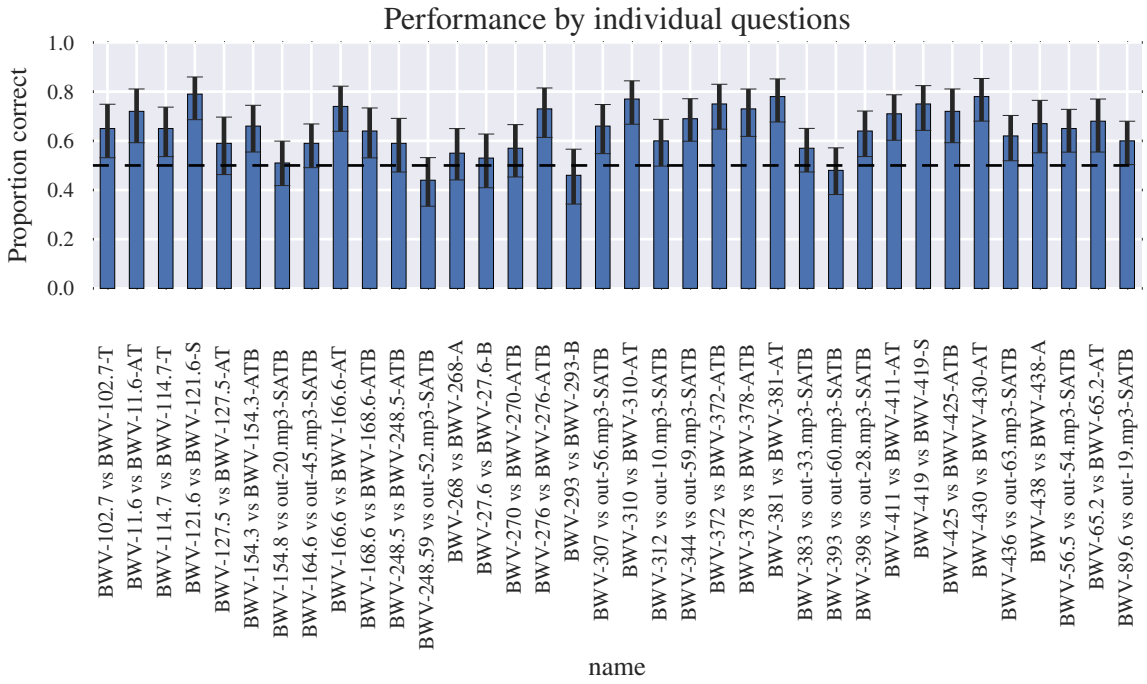



Fig. 1.9 Proportion of correct responses broken down by individual questions.

> Payments on mTurk are suggested to follow a reasonable hourly rate, with an example of $8 per hour or about 13c per minute. In practice, many mTurk tasks pay much less overall, with the median study paying just 5-10c for a task taking "a few minutes," like watching and providing feedback on 3 short (15-second) videos, summarizing a website, and evaluating hypothetical and real market products. Indeed, "wages" this low have been shown to result in lower quality output than could be had for no payment at all, by pure volunteers.

[3]

Paid service providers cost anywhere from $20 to $55 per month just for authoring tools and server space[7] At the time of writing, paid responses cost $1.50–$3.00 on SurveyMonkey [uks].

# References

[uks] Online research panel pricing | surveymonkey audience.

[2] Ariza, C. (2009). The interrogator as critic: The turing test and the evaluation of generative music systems. *Computer Music Journal*, 33(2):48–70.

[3] Downs, J. S., Holbrook, M. B., Sheng, S., and Cranor, L. F. (2010). Are your participants gaming the system?: screening mechanical turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2399–2402. ACM.

[4] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680.

[5] Pearce, M., Meredith, D., and Wiggins, G. (2002). Motivations and methodologies for automation of the compositional process. *Musicae Scientiae*, 6(2):119–147.

[6] Pearce, M. and Wiggins, G. (2001). Towards a framework for the evaluation of machine compositions. In *Proceedings of the AISB'01 Symposium on Artificial Intelligence and Creativity in the Arts and Sciences*, pages 22–32. Citeseer.

[7] Wright, K. B. (2005). Researching internet-based populations: Advantages and disadvantages of online survey research, online questionnaire authoring software packages, and web survey services. *Journal of Computer-Mediated Communication*, 10(3):00–00.