# Supplementary Online Material (SOM)

## Detailed ANN Results

### ANN model of the Participant AB:

The performance of ANN model of the Participant AB can be seen summarised below in Table 0.1.

*Table 0.1. Categorisation performance of the ANN model of participant AB and the corresponding human scores. Numbers in bold signify the highest confidence score on a given test.*



There were 7 correct categorisations on the 20 tests. The Bach category had the highest mean confidence score of 0.72 and the highest proportion of correct answers (5/5). The corresponding mean confidence scores for the other composers were 0.30 for Beethoven, 0.12 for Schubert and 0.24 for Mozart. The proportion of correct categorisations was 2/5 for Beethoven, 2/5 for Mozart and 0/5 for Schubert.

The human participant prediction pairs were directly compared to the top two predictions of the model – summarised in the Table 0.2 below.

*Table 0.2. The “per prediction pair” comparison metrics of the Participant AB and the corresponding ANN model.*



The comparison metrics were the following: 4/20 human and model prediction pairs were identical; 4/20 pairs matched disregarding the order of the predictions; 11/20 of the top predictions matched; 14/20 model’s prediction pairs contained the top human prediction; 20/20 model’s prediction pairs contained at least one composer that was also predicted by the human.

### ANN model of the Participant BC:

The performance of ANN model of the Participant BC can be seen summarised below in Table 0.3.

*Table 0.3. Categorisation performance of the ANN model of participant BC and the corresponding human scores. Numbers in bold signify the highest confidence score on a given test.*



There were 12 correct categorisations on the 21 tests. The Bach category had the highest mean confidence score of 0.70 and the highest proportion of correct answers (5/5). The corresponding mean confidence scores for the other composers were 0.42 for Beethoven, 0.18 for Schubert and 0.62 for Mozart. The proportion of correct categorisations was 2/5 for Beethoven, 4/5 for Mozart and 1/6 for Schubert.

The human participant prediction pairs were directly compared to the top two predictions of the model – summarised in the Table 0.4 below.

*Table 0.4. The “per prediction pair” comparison metrics of the Participant BC and the corresponding ANN model.*



The comparison metrics were the following: 1/21 human and model prediction pairs were identical; 2/21 pairs matched disregarding the order of the predictions; 11/21 of the top predictions matched; 14/21 model’s prediction pairs contained the top human prediction; 20/21 model’s prediction pairs contained at least one composer that was also predicted by the human.

### ANN model of the Participant CD:

The performance of ANN model of the Participant CD can be seen summarised below in Table 0.5.

*Table 0.5. Categorisation performance of the ANN model of participant CD and the corresponding human scores. Numbers in bold signify the highest confidence score on a given test.*



There were 8 correct categorisations on the 21 tests. The Mozart category had the highest mean confidence score of 0.51 and the highest proportion of correct answers (5/6). The corresponding mean confidence scores for the other composers were 0.00 for Beethoven, 0.10 for Schubert and 0.52 for Bach. The proportion of correct categorisations was 0/4 for Beethoven, 3/5 for Bach and 1/6 for Schubert.

The human participant prediction pairs were directly compared to the top two predictions of the model – summarised in the Table 0.6 below.

*Table 0.6. The “per prediction pair” comparison metrics of the Participant CD and the corresponding ANN model.*



The comparison metrics were the following: 0/21 human and model prediction pairs were identical; 0/21 pairs matched disregarding the order of the predictions; 7/21 of the top predictions matched; 12/21 model’s prediction pairs contained the top human prediction; 15/21 model’s prediction pairs contained at least one composer that was also predicted by the human.

### ANN model of the Participant DE:

The performance of ANN model of the Participant DE can be seen summarised below in Table 0.7.

*Table 0.7. Categorisation performance of the ANN model of participant DE and the corresponding human scores. Numbers in bold signify the highest confidence score on a given test.*



There were 13 correct categorisations on the 21 tests. The Bach category had the highest mean confidence score of 0.70 and the highest proportion of correct answers (5/5). The corresponding mean confidence scores for the other composers were 0.40 for Beethoven, 0.50 for Haydn and 0.48 for Mozart. The proportion of correct categorisations was 2/5 for Beethoven, 3/5 for Mozart and 3/5 for Haydn.

The human participant prediction pairs were directly compared to the top two predictions of the model – summarised in the Table 0.8 below.

*Table 0.8. The comparison metrics of the Participant DE and the corresponding ANN model.*



The comparison metrics were the following: 3/20 human and model prediction pairs were identical; 9/20 pairs matched disregarding the order of the predictions; 9/20 of the top predictions matched; 16/20 model’s prediction pairs contained the top human prediction; 20/20 model’s prediction pairs contained at least one composer that was also predicted by the human.

### ANN model of the Participant EF:

The performance of ANN model of the Participant EF can be seen summarised below in Table 0.9.

*Table 0.9. Categorisation performance of the ANN model of participant EF and the corresponding human scores. Numbers in bold signify the highest confidence score on a given test.*



There were 9 correct categorisations on the 20 tests. The Bach category had the highest mean confidence score of 0.66 and the highest proportion of correct answers (5/5). The corresponding mean confidence scores for the other composers were 0.46 for Beethoven, 0.20 for Mozart and 0.00 for Schubert. The proportion of correct categorisations was 3/5 for Beethoven, 1/5 for Mozart and 0/5 for Schubert.

The human participant prediction pairs were directly compared to the top two predictions of the model – summarised in the Table 0.10 below.

*Table 0.10. The comparison metrics of the Participant EF and the corresponding ANN model.*



The comparison metrics were the following: 2/20 human and model prediction pairs were identical; 2/20 pairs matched disregarding the order of the predictions; 8/20 of the top predictions matched; 11/20 model’s prediction pairs contained the top human prediction; 18/20 model’s prediction pairs contained at least one composer that was also predicted by the human.

### ANN model of the Participant FG:

The performance of ANN model of the Participant FG can be seen summarised below in Table 0.11.

*Table 0.11. Categorisation performance of the ANN model of participant FG and the corresponding human scores. Numbers in bold signify the highest confidence score on a given test.*



There were 12 correct categorisations on the 21 tests. The Beethoven category had the highest mean confidence score of 0.84 and the highest proportion of correct answers (5/5). The corresponding mean confidence scores for the other composers were 0.52 for Haydn, 0.50 for Mozart and 0.04 for Schubert. The proportion of correct categorisations was 3/5 for Haydn, 4/6 for Mozart and 0/5 for Schubert.

The human participant prediction pairs were directly compared to the top two predictions of the model – summarised in the Table 0.12 below.

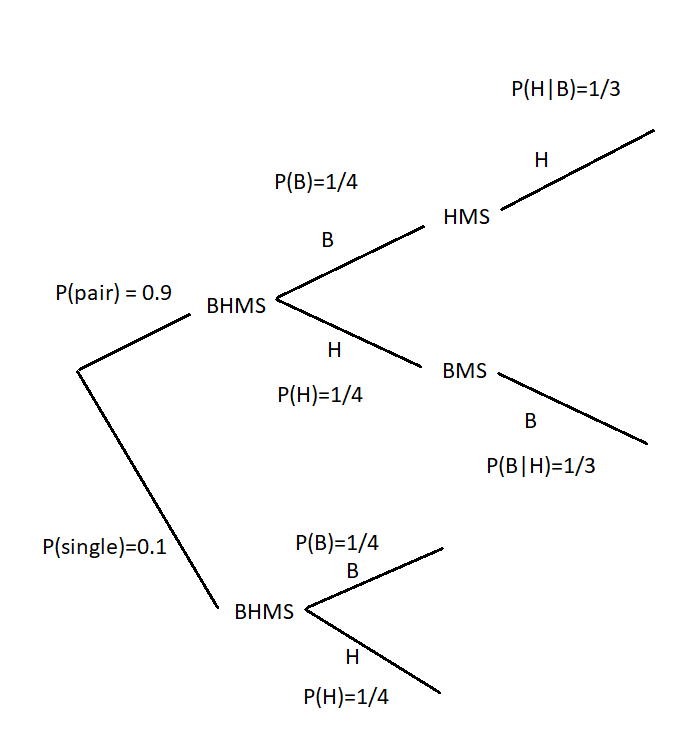
*Table 0.12. The “per prediction pair” comparison metrics of the Participant FG and the corresponding ANN model.*



The comparison metrics were the following: 1/21 human and model prediction pairs were identical; 2/21 pairs matched disregarding the order of the predictions; 5/21 of the top predictions matched; 11/20 model’s prediction pairs contained the top human prediction; 17/21 model’s prediction pairs contained at least one composer that was also predicted by the human.

## SOM 1: Calculation of statistical significance for CogAct simulation of human data

To establish the statistical significance of the CogAct (and ANN) simulations, we need to answer the following question: what is the probability of matching the participants’ music categorisation answers by chance?



*Figure A1. The probability tree of picking a matching answer.*

I note the two possible outcomes: one, that human participant may not be fully certain and pick her answer as a pair of composers (i.e., (first choice, second choice)); second, she may be 100% certain in her answer and thus pick a single composer. I will now determine the probabilities of picking an answer that matches that of a human by chance.

Modelling the paired answers

Firstly, I will consider the paired answers. Let us assume that participant was choosing among 4 composers B (Bach), H (Haydn), M (Mozart), S (Schubert) and that her choice was (B, H).

The probability of picking a paired answer is calculated from the structure of the Likert scale used for both human and model scores. Confidence of 1 (i.e., 100%) led to a single answer (e.g., “Bach”), and all other confidences implied paired answers with a first and a second preference (e.g., “Bach, “Haydn”).

*P(pair) = 0.9*

*P(single) =0.1*

Then, the probability of picking an *identical* pair (i.e., the “Identical” metric) is the following:

*P(“Identical” pair) = P(pair)\*P(B&H) = P(pair)\*P(B)\*P(H|B) = 9/10 \* 1/4 \* 1/3 = 3/40*

The probability of picking a pair of composers where order is not important (the “Both match” metric) can be calculated using the following:

First, the probability of picking the reverse order (H&B) is:

*P(pair)\*P(H&B) = P(pair)\*P(H)\*P(B|H) = 9/10 \* 1/4 \* 1/3 = 3/40*

Thus,

*P(“Both match” pair) = P(pair)\*(P(B&H) or P(H&B)) = P(pair)\*P(B&H) + P(pair)\*P(H&B) = 3/40 + 3/40 = 0.15*

The probability of picking an answer where its top prediction matches human’s top prediction (i.e., the “Tops match” metric) has the following calculation:

*P(“Tops match” answer) = P(pair)\*(P(B&H) + P(B&M) + P(B&S)) + P(single)\*P(B) = 9/10 \* 3/12 + 1/10 \* 1/4 = 9/40 + 1/40 = 10/40 = 1/4*

The probability of picking an answer where at least one pick matches human’s top pick (i.e., the “One matches top” metric) has the following calculation:

*P(“One matches top” pair) = P(pair)\*(P(B&H) + P(H&B) + P(B&M) +P(M&B) +P(B&S) +P(S&B)) + P(single)\*P(B) = 9/10\*6\*1/12 +1/10\*1/4 = 6\*3/40 + 1/40 = 19/40 = 0.475*

The probability of picking an answer were at least one composer matches at least one of human’s picks (i.e., the “Single match” metric) is the following:

*P(“Single match” pair) = P(pair)\*(1 – (P(S&M) + P(M&S))) + P(single)\*(P(B) +P(H)) = 30/40 + 1/10\*2/4 = 32/40 = 0.8*

(See Figure A1 for the diagram of the described probabilities.)

Modelling the single answers

Now, we can look at the probabilities of the model matching the single human answer. Let us assume that participant was choosing among 4 composers B (Bach), H (Haydn), M (Mozart), S (Schubert) and that her choice was (B). Then, the probability of picking the *identical* answer is the following:

*P(“Identical” single) = P(single)\*P(B) = 1/10\*1/4 = 1/40*

The probability of picking an answer where the order is not important is the same as the above:

*P(“Both match” single) = P(single)\*P(B) = 1/40*

The probability of picking an answer where its top prediction matches human’s only prediction (i.e., the “Tops match” metric) has the following calculation:

*P(“Tops match” single) = P(pair)\*(P(B&H) + P(B&M) + P(B&S)) + P(single)\*P(B) = 3\*3/40 + 1/40 = 10/40 = 1/4*

The probability of picking an answer where at least one pick matches human’s top pick (i.e., the “One matches top” metric) has the following calculation:

*P(“One matches top” single) = P(pair)\*(P(B&H) + P(H&B) + P(B&M) +P(M&B) +P(B&S) +P(S&B)) + P(single)\*P(B) = 6\*3/40 + 1/40 = 19/40 = 0.475*

The probability of picking an answer were at least one composer matches at least one of human’s picks (i.e., the “Single match” metric) is the following:

*P(“Single match” single) = P(pair)\*( P(B&H) + P(H&B) + P(B&M) +P(M&B) +P(B&S) +P(S&B)) + P(single)\*P(B) = 6\*3/40 + 1/40 = 19/40 = 0.475*

(See Figure A1 for the diagram of the described probabilities.)

Estimating the general outcome on a single trial

Having calculated the probabilities for paired and single answers, we can now produce the general/combined probabilities for each of the metrics. There were 122 predictions for both the human participants and the corresponding models. Humans had 82 predictions in the form of pairs where they were less than certain about the answer (e.g., “this piece is likely by Bach, but it could also be Haydn”) and 40 predictions where they were 100% certain and gave a single answer (e.g., “it just has to be Bach).

Thus, the general probability of randomly picking identical answers was calculated in the following way:

*P(“Identical”) = 82/122 \* P(“Identical” pair) + 40/122 \* P(“Identical” single) = (82\*3/40 +40\*1/40)/122 = 0.0586*

*P(“Both match”) = (82 \* P(“Both match” pair) + 40 \* P(“Both match” single))/122 = (82\*0.15 + 40\*1/40)/122 = 0.1090*

*P(“Tops match”) = (82 \* P(“Tops match” pair) + 40 \* P(“Tops match” single))/122 = (82\*1/4 + 40\*1/4)/122= 0.25*

*P(“One matches top”) = (82 \* P(“One matches top” pair) + 40 \* P(“One matches top single”)) /122 = (82 \* 0.475 + 40 \* 0.475) /122 = 0.475*

*P(“Single match”) = (82 \* P(“single match” pair) + 40 \* P(“Single match” single)) /122 = (82\*0.8 + 40\*1/4) /122 = 0.6934*

Computing statistical significance of CogAct models

With the base level probabilities in place, I will now compute the p-values for each of the 5 metrics across all the participants. This would estimate the probability of false positive results.

The cumulative/total metric scores *for Identical, Both match, Tops match, One matches top* and *Single Match* were 15, 27, 48, 77 and 107 respectively.

total = [15,27,48,77,107]

With the above information in hand, the Bernoulli formula for discrete binomial distribution was used to estimate the statistical significance of the results. The p-values of the metrics were calculated as follows:

Here, *p* is the probability of the corresponding metric getting a match with the human score, *n* is the number of trials and *k* is the number of matching answers.

The computation was performed with the following code:

**def** nCk(n, k): *#calculating “n choose k” combinations*

n\_choose\_k = math.factorial(n) / (math.factorial(k) \* math.factorial(n - k))

**return** n\_choose\_k

**def** prob\_calc(n, k, prob): *#i.e., (trials, hits, prob\_of\_matches)*

total\_p = prob\*\*(n-(n-k)) \* (1-prob)\*\*(n-k) \*nCk(n,k)

**return** total\_p

**def** at\_least\_n(n, k, prob):

total\_p = 0

**while** k <= n:

total\_p += prob\_calc(n, k, prob)

k+=1

**return** total\_p

*"""AB-> 6 ones / 20;*

*BC -> 7 ones /21;*

*CD -> 10 ones /21 (not incl Schubert) 15 ones including Schubert;*

*DE -> 0 ones /20;*

*EF -> 8 ones /20;*

*FG -> 4 ones /20;*

*ALL\_subjects -> 40 /122 #40 '1' ratings on 122 tests"""*

ab = [4,4,9,12,18]

bc = [1,2,9,14,20]

cd = [0,2,8,11,13]

de = [5,10,8,15,20]

ef = [3,3,9,11,18]

fg = [2,6,5,14,18]

all\_metrics = [ab,bc,cd,de,ef,fg]

total\_metrics = [0,0,0,0,0] *#add all cols*

**for** m **in** all\_metrics:

**for** i **in** range(len(m)):

total\_metrics[i] += m[i]

print(total\_metrics)

t\_id = (40 \* 1/40 + 82 \* 3/40) /122 *#general prob of Identical, combining Identical pair and Identical single probs*

t\_bm = (82 \* 0.15 + 40 \* 1/40) /122

t\_tm = 1/4

t\_omt = (82\*0.475 +40\*0.475)/122

t\_sm = (82 \* 0.8 + 40 \* 0.475)/122

p\_sm = at\_least\_n(122,107, t\_sm) *#Prob of at least 107 Single Match*

p\_omt = at\_least\_n(122, 77, t\_omt) *#Prob of at least 77 One Matches Top*

p\_tm = at\_least\_n(122, 48, t\_tm) *#Prob of at least 48 Tops Match*

p\_bm = at\_least\_n(122, 27, t\_bm) *#Prob of at least 27 Both Match*

p\_id = at\_least\_n(122, 15, t\_id) *#Prob of at least 15 Identical*

P(at least 15 “Identical” answers) = 0.00529

P(at least 27 “Both match” answers) = 0.0002617

P(at least 48 “Tops match” answers) = 0.0003373

P(at least 77 “One matches top” answers) = 0.0003734

P(at least 107 “Single match’ answers) = 1.709\*e^-6

To adjust for multiple hypothesis testing, a Bonferroni adjustment was made. Since there were 5 hypotheses and the confidence threshold was 0.05, the adjusted confidence threshold became 0.05 / 5 = 0.01. Since the p-values above fall into the 0.01 threshold we may now reject the null hypothesis and accept the experimental hypothesis: CogAct models of music concept formation produce good fit to human data.

## SOM 2: Calculating statistical significance for ANN simulation of human data

This appendix reports calculating statistical significance for ANN simulation of human data.

See Appendix 8 for calculating general probabilities of matching a human answer by chance.

The cumulative/total ANN metric scores *for Identical, Both match, Tops match, One matches top* and *Single Match* were 11, 19, 51, 78 and 110 respectively.

The Bernoulli formula for discrete binomial distribution was used to estimate the statistical significance of the results. The p-values of the metrics were calculated as follows:

Here, *p* is the probability of the corresponding metric getting a match with the human score, *n* is the number of trials and *k* is the number of matching answers. N.B.: the code that was used to perform the calculation is presented in Appendix 8 above.

P(at least 11 “Identical” answers) = 0.1030

P(at least 19 “Both match” answers) = 0.071

P(at least 51 “Tops match” answers) = 3.576\*e^-5

P(at least 78 “One matches top” answers) = 0.000188

P(at least 110 “Single match’ answers) = 3.807\*e^-8

To adjust for multiple hypothesis testing, a Bonferroni adjustment was made. Since there were 5 hypotheses and the confidence threshold was 0.05, the adjusted confidence threshold became 0.05 / 5 = 0.01. Since the 3 of the 5 metrics’ p-values above fall into the 0.01 threshold we may now cautiously reject the null hypothesis and accept the experimental hypothesis: ANN models of music concept formation produce good fit to human data.