**Hypothesis-specific, confirmatory analysis plan**

All analysis will be implemented in R 3.3.2.

**Omnibus test**

Using the *lme4* package, a mixed effects multiple linear regression model will be fitted as per the study design with maximal random effects (Barr, Levy, Scheepers, & Tily, 2013).

Two models will be fit: one for prediction and one for precision (certainty ratings for humans), where these are the predicted values. Both models will include *Time* (Before, After) and *Type* (Human, IDyOM) as fixed effects, with interactions modelled. Random effects include random intercepts on participants and random slopes on melody. IDyOM information content and entropy will be calculated using the linked pitch and pitch interval viewpoints for this omnibus analysis as well as for Analysis 2a); all three viewpoints (pitch, pitch interval, scale degree) and their combinations will be investigated in Analysis 2b).

*prediction.model <- lmer(prediction ~ Time \* Type + (1+melody.id|Participant), data = tidydata)*

*precision.model <- lmer(precision ~ Time \* Type + (1+melody.id|Participant), data = tidydata)*

A significant effect of Time with a negative coefficient would indicate a decrease in prediction and precision from “before” to “after”; a significant effect of Type with a negative coefficient would indicate that the computational model has lower prediction and precision values than humans; a significant interaction between these fixed effects would indicate that humans and the computational model change their behaviour differently from “before” to “after”.

Where relevant to each hypothesis, follow-up t-tests with Bonferroni correction (p = .05/4 = .0125) are described below in addition to tests specifically related to each hypothesis.

**Hypothesis 1a)** Prediction and precision will be poor for listeners who are unfamiliar with serial music.

**Analysis 1a)**

One-sided t-test, where

x = prediction ratings from the “Before” test

mu = 5 (1 point above the middle of the rating scale, where 1 = low surprise and 7 = high surprise)

alternative = “greater”

*t.test(prediction.ratings.before, mu = 5, alternative = “greater”)*

One-sided t-test, where

X = certainty ratings from the “Before” test

Mu = 5 (1 point above the middle of the rating scale, where data has been transformed so that 1 = completely sure and 7 = completely unsure)

Alternative = “less”

*t.test(certainty.ratings.before, mu = 5, alternative = “greater”)*

**Hypothesis 1b)** Expected uncertainty predicts that prediction will improve but precision will remain poor, after listeners have formally studied serial music.

**Analysis 1b)**

Paired t-test, where

X = prediction ratings from the “Before” test

Y = prediction ratings from the “After” test

Alternative = “greater”

*t.test(prediction.ratings.before, prediction.ratings.after, alternative = “greater”)*

Paired t-test, where

X = certainty ratings from the “Before” test

Y = certainty ratings from the “After” test

Alternative = “two-sided” (default)

*t.test(certainty.ratings.before, certainty.ratings.after)*

Equivalence test (using the TOSTER package), where

“Before” and “After” prediction ratings are compared and the upper equivalence bound is set to .5 in raw units, corresponding to a change in half a rating point.

*dataTOSTpaired(prediction.ratings, pairs = list(c(i1 = “prediction.ratings.before”, i2 = “prediction.ratings.after”)), high\_eqbound = .5, eqbound\_type = “raw”)*

Equivalence test (using the TOSTER package), where

“Before” and “After” certainty ratings are compared and the upper and lower equivalence bounds are set to .5 in raw units.

*dataTOSTpaired(certainty.ratings, pairs = list(c(i1 = “certainty.ratings.before”, i2 = “certainty.ratings.after”)), low\_eqbound = -.5, high\_eqbound = .5, eqbound\_type = “raw”)*

**Hypothesis 2a)** A computational model (IDyOM; Pearce, 2005; 2018) that has been exposed to serial music will predict serial music more accurately and with better certainty than humans

**Analysis 2a)**

Between-samples t-test, where

X = Human prediction ratings from the “After” test (scaled)

Y = IDyOM information content (linked pitch and interval viewpoints) from the “After” test (scaled)

Alternative = “less”

*t.test(prediction.ratings.after, ic.after, alternative = “less”)*

Between-samples t-test, where

X = Human precision ratings from the “After” test (values subtracted from 7 and scaled)

Y = IDyOM entropy (linked pitch and interval viewpoints) from the “After” test (scaled)

Alternative = “less”

*t.test(precision.ratings.after, entropy.after, alternative = “less”)*

**Hypothesis 2b)** Information content and entropy produced by the computational model using the linked musical properties pitch and pitch interval will correlate with human ratings of prediction and precision better than measures produced using these properties individually or including the property scale degree.

**Analysis 2b)**

Pearson’s correlation coefficient R will be calculated between human ratings of prediction and IDyOM information content produced by each combination of musical properties (viewpoints). The same will be done for human ratings of precision and IDyOM entropy. These tests will be performed on pooled “Before” and “After” data and on “Before” and “After” data separately.

*viewpoints <- c(pitch, interval, scale degree, pitch-interval, pitch-scale degree, interval-scale degree, pitch-interval-scale degree)*

*for(i in 1:length(viewpoints)){*

*cor.test(prediction.ratings, ic.viewpoint[i])*

*}*