Crossmodality Toolkit Paper - Draft 1

Alan Nielsen, Justin Sulik, Bill Thompson, Hannah Little, Yasamin Motamedi, Kevin Stadler, Ashley Micklos, Piera Filippi, Marcus Perlman, Andrea Ravignani, Mark Dingemanse, & Stephen Levinson

November 2, 2017

##### This is a line break

##### This is a line break

Affiliatons: Max Planck Institute for Psycholinguistics ...

##### This is a line break

##### This is a line break

Correspondence: Alan Nielsen, Max Planck Institute for Psycholinguistics, Nijmegen, 6512 HK, Netherlands. Email: [alan@languageevolution.com](mailto:%20alan@languageevolution.com)

##### This is a line break

##### This is a line break

##### This is a line break

Manuscript word count: XXXX

###### This is a Page Break

# ABSTRACT

This is where the abstract goes. Write it.

### Highlights:

This is where highlights go (if whatever journal we submit to uses them)

### Keywords:

crossmodality; synesthesia; etc; etc

###### This is a Page Break

# Introduction

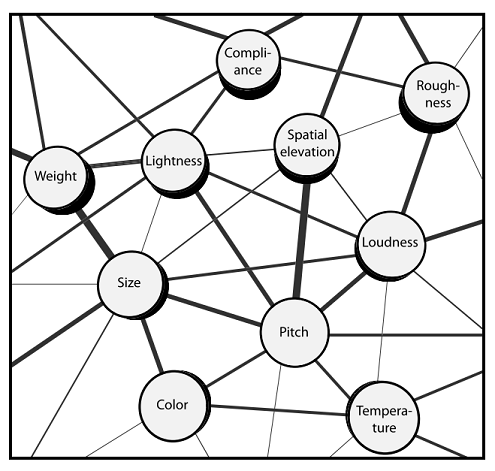
Humans are not unbiased perceivers of the world, taking in information from the environment and processing it in a vacuum. For example, most people associate high-pitched sounds with small, fast moving objects that are located high in space. Similarly, despite 'maluma' not being a word in English, experimental participants generally agree that it is a better label for an unfamiliar curvy object than a jagged one. These types of biases have been documented since at least the late 19th century in the cognitive sciences literature (e.g. Kohler, 1929), and have an even longer pedigree in philosophy and the humanities (Plato ref). Because of their long history, these biases have gone by a number of names (phonetic symbolism: Sapir, 1929; sound symbolism: REF; metaphorical mappings: Evans & Treisman, 2010) depending on the field of study and particular interests of the researcher in question. Recently, however, the term *crossmodal correspondences* has become most common (Parise, 2016; Spence, 2011), and researchers in various fields have suggested that these crossmodal correspondences are fundamental features of human cognition (Deroy & Spence, 2016) that might have important downstream effects on the structure of languages or other cultural traditions (Cuskley thesis ref; Nielsen, 2017; other refs?).

In the past, and increasingly in the last decade, researcher have compiled a substantial catalogue of these crossmodal correspondences, using a variety of stimuli and experimental paradigms. On the stimulus front, researchers have made use of materials that varied enormously in terms of their complexity - from pure tones (Tarte, 1981) and colored triangles (Tarte, 1974) to complex musical scores (Palmer et al., 2013) and videos of dance performances (Aronoff, 2006). Similarly, evidence for putatively crossmodal correspondences has been recorded using methodologies from explicit matching tasks (Sapir, 1929) to more implicit psychophical discrimination (Parise and Spence, 2009) or association (Parise and Spence, 2012) tasks (reviewed in Parise, 2016).

As noted by others (Spence, 2011; Parise, 2016), it seems to be the case that wherever we look for evidence of a crossmodal association, we find it, so what are we to make of the growing crossmodal correspondence literature, especially with regards to what it tells us about cognition and its effects on language and related processes? One possibility is that it is indeed the case that crossmodal correspondences can be (and are) manifest in all possible configurations, but there are a number of (non-exclusive) alternatives:

1. **Researcher Bias-** *Researchers may be biased in the types of correspondences that they are testing by the fact that they, themselves, make those associations.*
2. **The File Drawer Problem-** *Evidence for crossmodal correspondences might be inflated by the file-drawer problem, with non-significant results left unpublished.*
3. **Task Demands-** *A wide variety of tasks are used to query crossmodal correspondences. The more explicit of these tasks might actually query response strategies that would have no impact in more implict tests.*
4. **Stimulus Selection-** *Stimuli used in typical crossmodal correspondence tasks are generally not shared between studies, do not vary parametrically, and are presented as binary pairs (e.g. a "high" vs "low" pitched tone) which assume that crossmodal correspondences are monotonic (Parise, 2016).*

If we ignore these possibilities, we can construct a network of crossmodal correspondences (Figure 1, below):



**Figure 1-** *A network of a sample of observed crossmodal correspondences, from Parise (2016)*

This type of network representation is useful, and does indeed catalogue a subset of the multitude of crossmodal correspondences, but comparing the associations to one another is problematic, given different stimuli, experimental protocols, and standards of evidence. This is further complicated by the fact that crossmodal correspondences are not likely to represent a unitary, ungraded, monotonic phenomenon (Westbury, 2005), but to have multiple causes, including learning and mediation (Sidhu & Pexman, 2017).

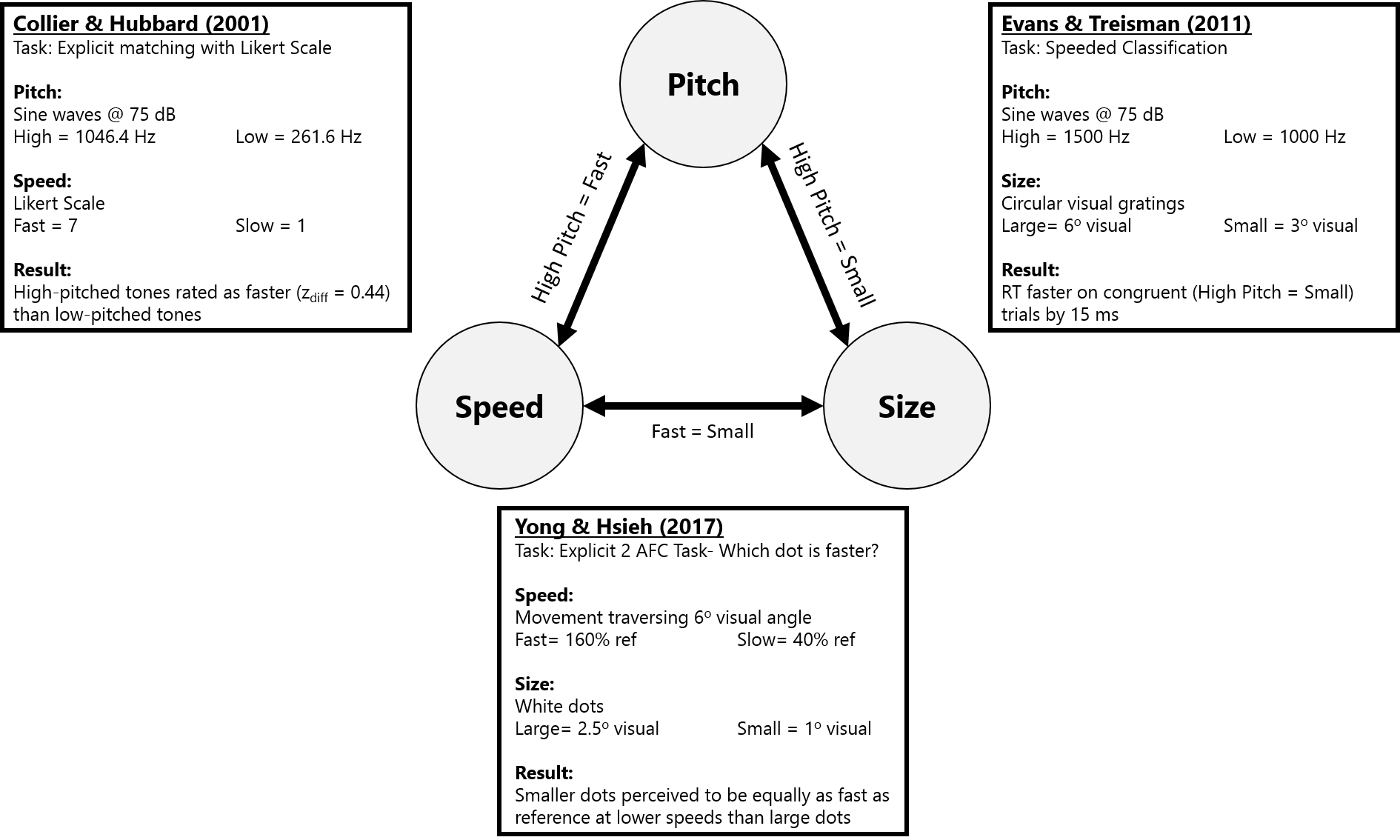
Resolving these issues will be problematic - and a major focus of multiple research streams in the coming years. In the interim, however, there are improvements to be made that can begin to untangle the gordian knot of crossmodal correspondences. Here, we present the results of a set of experiments that attempts to serve as a starting-point for further systematic exploration of crossmodal correspondences. To that end, we conducted a large-scale experiment exhaustively testing a set of crossmodal correspondences between nine stimulus domains using shared stimuli and testing multiple correspondences with each participant. This manipulation allows us to, for the first time (to our knowledge), collect and analyse a fully-specified netword of crossmodal correspondences.

##### This is a line break

# Experiments

Papers exploring crossmodal correspondences often focus on a single association, such as the well-studied Takete-Maluma effect (Kohler, 1947), where plosive consonants like */k/* and */t/* are associated with jagged visual contours and more mellifluous sonorant consonants like */m/* and */l/* with curvier images (Nielsen & Rendall, 2011; cf. Styles & Gawne, 2017 for a recent review). This approach, of course, has its strengths, especially as researchers explore different methodologies and center in on increasing accurate descriptions of behaviour and its mechanistic underpinnings (e.g. Jones et al. ref). This focus on single correspondences, however, makes comparing the results of various experiments difficult. Consider Figure 2, below:

##### This is a line break



**Figure 2-** *Crossmodal correspondences between Pitch, Size, and Speed, from three separate studies demonstrates the difficulty in relating experimental findings that use different stimuli and designs*

##### This is a line break

Figure 2 demonstrates that even with all of the pertinent information from a set of crossmodal correspondence experiments, it can be difficult to make meaningful predictions to new stimuli or new experimental paradigms. In a novel experiment, how might pure tones at 1500 Hz (high) and 1000 Hz (low; Evans & Treisman, 2011) be matched to white dots moving at either 40% (slow) or 160% (fast; Yong & Hsieh, 2017) the speed of a reference object? Would the results be similar to those obtained by Collier & Hubbard (2001) studying the "same" crossmodal correspondence? Would the nature of the task matter? What about the directionality of the association being tested (e.g. whether participants were choosing which of two moving images corresponds to a high-pitched sound, or decide whether a high vs. low-pitched sound corresponds to a fast-moving object). In the best cases, well-intentioned researchers have explored over a dozen crossmodal correspondences (e.g. Lindauer, 1990; other refs?), but even these types of studies do not typically make a sufficient impact that their stimuli are re-used or even replicated.

The primary goal of the work presented here is thus to move towards a standardized set of stimuli that can be used to exhaustively test a large number of combinations of crossmodal correspondences. To that end, we make use of a relatively straightforward, totally explicit 2-alternative-forced-choice experimental protocol that allows for easy online deployment and the collection of large amounts of data. We recognize that this explicit experimental protocol might not be ideal, and further that our stimuli are binary, rather than graded, and overall more simplistic than the standards of more detailed, focused experimental studies. Thus, it is not the intention of the data presented here to supplant the majority of previously established work, or to enshrine the stimuli or methodology as the most appropriate for the study of crossmodal correspondences. To the contrary, we hope that by presenting a coherent set of data we will motivate further work on crossmodal correspondences.

## Methods

Although we collected data for three separate experiments, the methodology and stimuli used for all experiments was identical. Thus, before describing the structure of each individual experiment we will discuss stimulus materials and the experimental interface.

### Participants

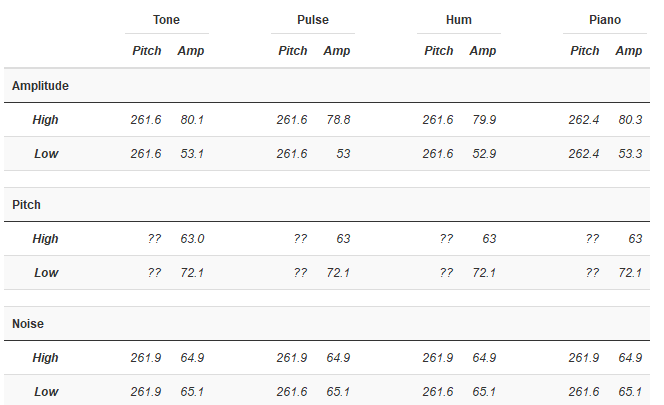
A total of 319 participants were recruited from Amazon Mechanical Turk for participation in our experiments. Of these, responses were collected from 61 participants as a Pilot. Subsequently, we collected data from an additional 144 participants in the main condition (Experiment 1A), 55 participants for Experiment 1B, and 59 participants for Experiment 1C. Participants in the Pilot and in Experiment 1A were paid 1.50D US for their participation, which took approximately 15 minutes. Participants in Experiment 1B and 1C were paid 1.00 USD for their participation, which took approximately 10 minutes. Ethical approval was obtained from the Max Planck Institute for Psycholinguistics, Ethics # PUTNUMBER HERE.

### Materials

Stimuli for the Experiments presented here were created to test for crossmodal associations between 9 domains: Emotion, Color, Brightness, Pitch, Noisiness (Spectral Density), Amplitude (Loudness), Speed, Size, and Shape. There stimulus domains were chosen by virtue of being well studied, and relatively easy to present to participants via an online interface. For each domain, four sets of stimulus tokens that varied only on a single dimension were created. The details of each set of stimuli can be found below, and all stimuli can be found online in the [github respository](https://github.com/hecticdialectic/CMT/tree/master/Experiment/Stimuli) for this paper.

#### Auditory Stimuli

Auditory stimuli were created from four sources: a pure tone, a pulse, a hum, and a piano note. Table 1, below, shows the acoustical values of the stimuli.



**Table 1-** *Analysis of the Acoustic Features of Auditory Stimuli for Experiment I*

##### This is a line break

BASIC DETAILS ABOUT GENERATION GO HERE- HOW WERE THEY GENERATED. WHAT PROGRAM. WHAT LENGTHS. ETC

For pairs of stimuli contrasting in Amplitude, we generated stimuli from each of the four sources (tone, pulse, hum, piano) with a frequency of 523.2 Hz (based on the middle frequency from Collier & Hubbard (2001)). High Amplitude tokens were generated at 80 dB, with Low Amplitude tokens at 53 dB (following Marks, 1989).

For pairs of stimuli contrasting in Pitch, we generated stimuli from the four sources, correcting amplitude based on equal loudness curves from Suzuki & Takeshima (ref). Thus, following Bien et al. (2012) High Pitch tokens were generated with a pitch of 4500 Hz at 63 dB, while Low Pitch tokens were generated with a pitch of 250 Hz at 72 dB.

For pairs of stimuli contrasting in their level of noise, Low Noise tokens were generated with a pitch of 523.2 Hz at 65 dB. High Noise tokens were based on low-noise tokens, but had noise overlaid DETAILS ABOUT THIS HERE.

#### Visual Stimuli

All visual stimuli were presented in 240 x 240 resolution. OTHER DETAILS JUSTIN?

##### This is a line break

*Brightness*

##### This is a line break



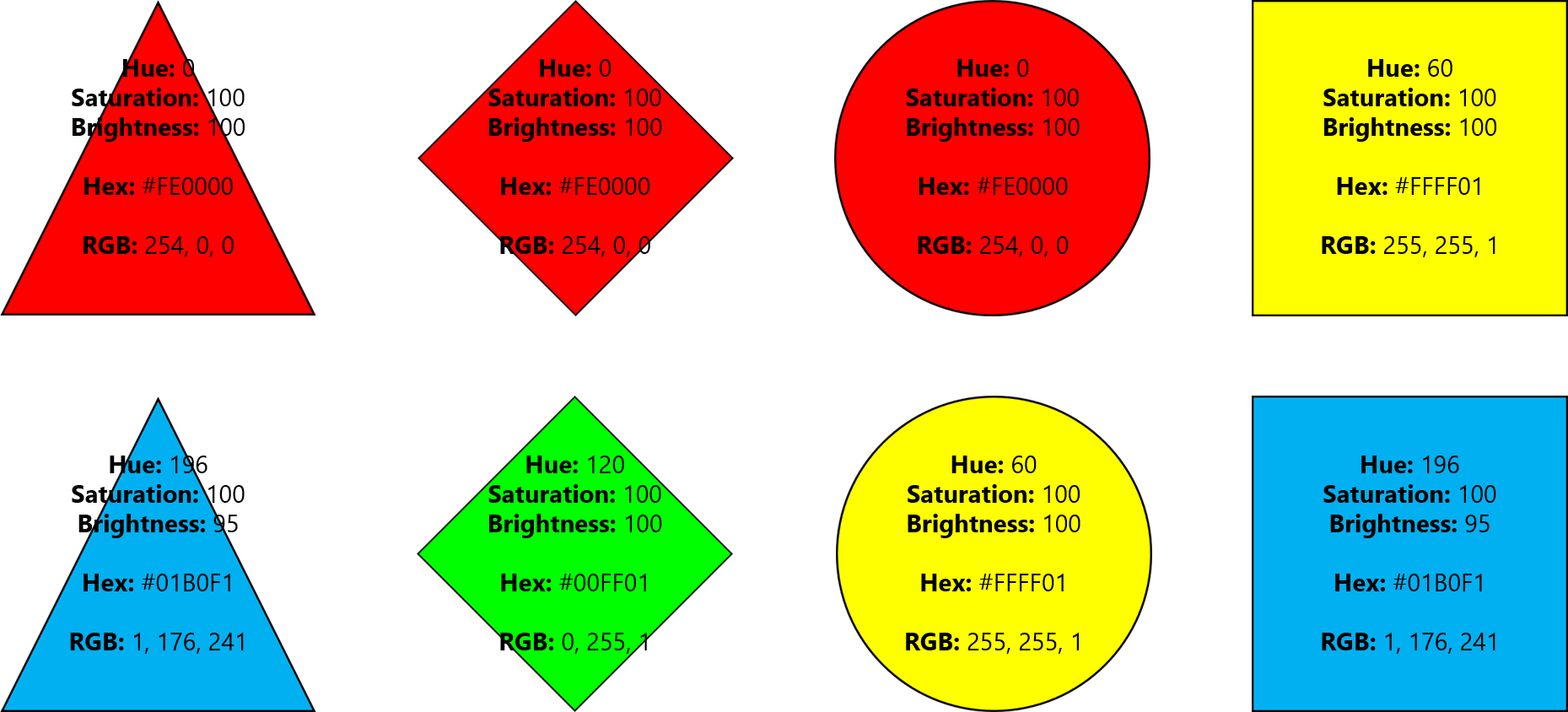
**Figure 3-** *Visual Stimuli showing pairs of High Brightness (top) and Low Brightness (Bottom) Shapes from Experiment 1. Inside of each figure you can see listed the HSB color values, Hex Representation, and RGB code for each color*

##### This is a line break

For pairs of stimuli contrasting in Brightness, we generated images of four types: Blue Squares, Yellow Circles, Red Triangles, and Green Diamonds. In each pair of images, hue and saturation were kept constant, with Brightness being the only difference between pairs, according to the HSB color model (REF).

*Color*

##### This is a line break



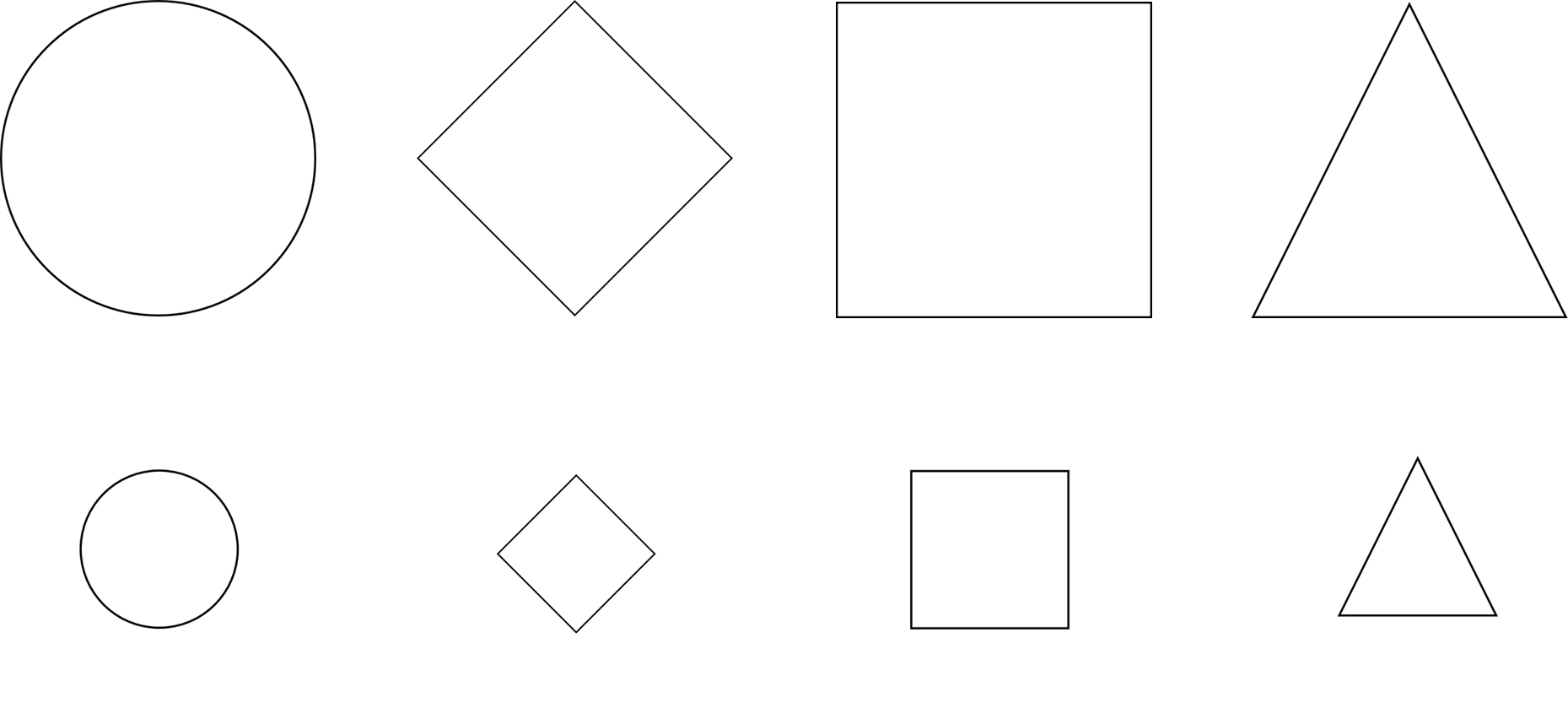
**Figure 4-** *Visual Stimuli showing pairs contrasting in color from Experiment 1. Inside of each figure you can see listed the HSB color values, Hex Representation, and RGB code for each color*

##### This is a line break

For pairs of stimuli contrasting in color, we generated four sets of images: triangles (red vs. blue), diamonds (red vs. green), circles (red vs. yellow), and squares (yellow vs. blue) that were of equal brightness according to HSB (Should we put in a note here discussiong perceptual brightness?), but differed in their hue. Because we used four different hues, we did not exhaustively test all six color comparisons, instead opting for the comparison of red to all other colors, with the fourth slot filled by the comparison of yellow to blue.

*Size*

##### This is a line break



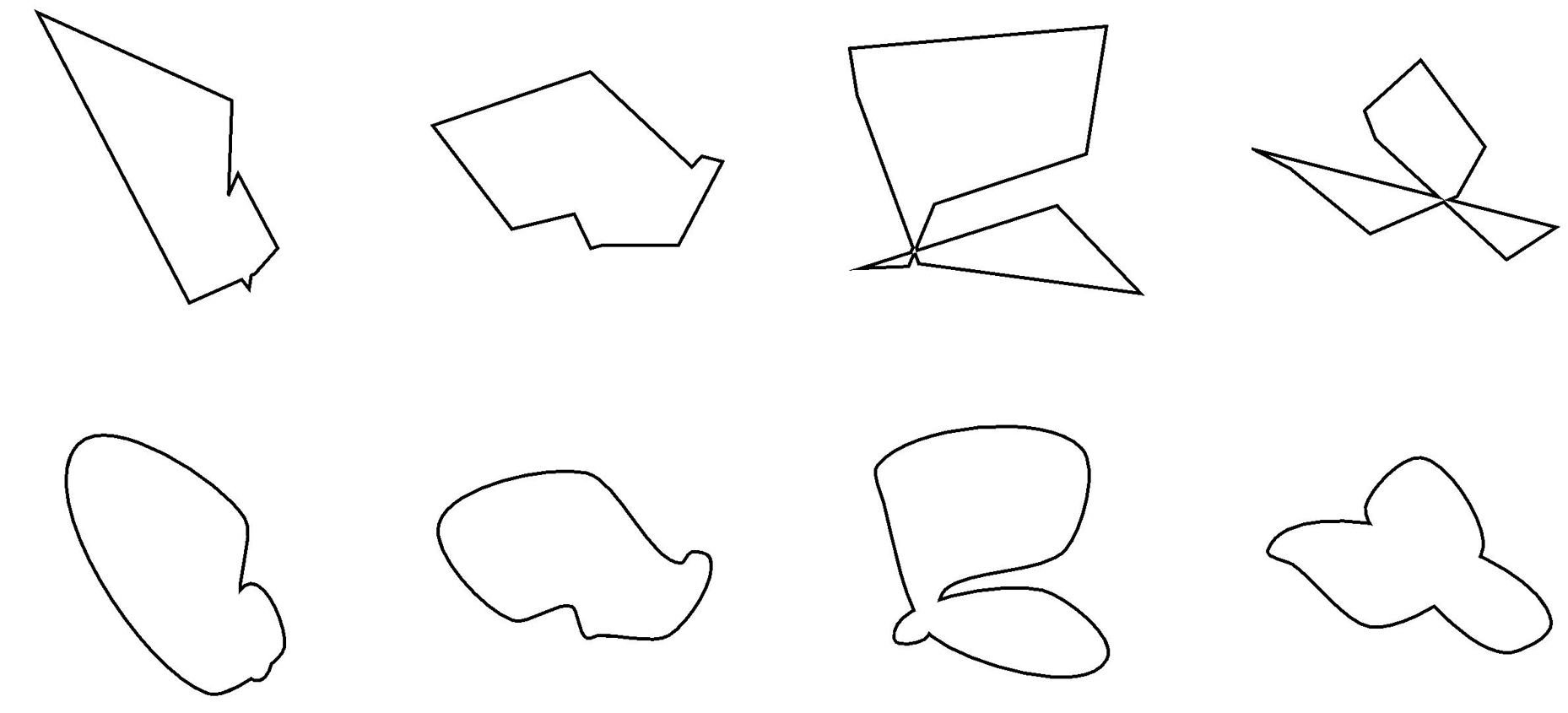
**Figure 5-** *Visual Stimuli showing pairs of Large (Top) and Small (Bottom) images from Experiment 1*

##### This is a line break

For pairs of stimuli contrasting in size, we generated four sets of black and white images: circles, diamonds, squares, and triangles. Large images were four times the size of their smaller counterparts. MORe DETAILS?

*Shape*

##### This is a line break



**Figure 6-** *Visual Stimuli showing pairs of Jagged (Top) and Curved (Bottom) images from Experiment 1*

##### This is a line break

For pairs of stimuli contrasting in shape, we generated four sets of black and white images that were either Jagged or Curved, following the procedure described in Nielsen & Rendall, 2011. Each image pair was generated from the same initial seed, and thus was overall of a similar shape and size.

*Speed*

##### This is a line break

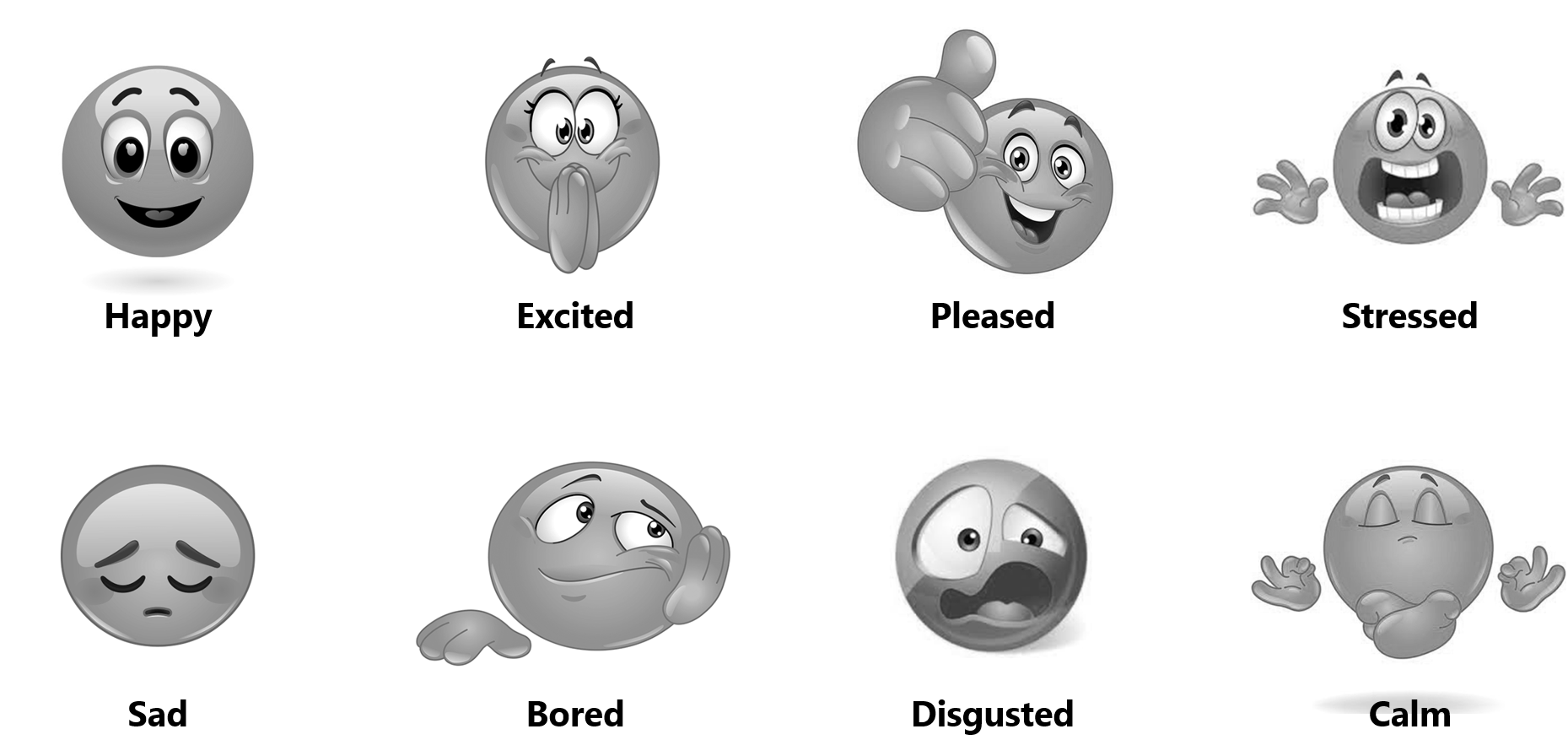
FIGURE GOES HERE?

##### This is a line break

For pairs of stimuli contrasting in speed, we created gifs of black squares moving backwards and forwards horizonally. The movement trajectories of these squares were generated from HANNAH PROVIDES DETAILS. Fast and Slow stimuli within a pair followed identical trajectories, but DETAILS ABOUT THE DIFFERENCES.

*Affect (Emotion)*

##### This is a line break



**Figure 7-** *Visual Stimuli showing pairs of Emotions from Experiment 1*

##### This is a line break

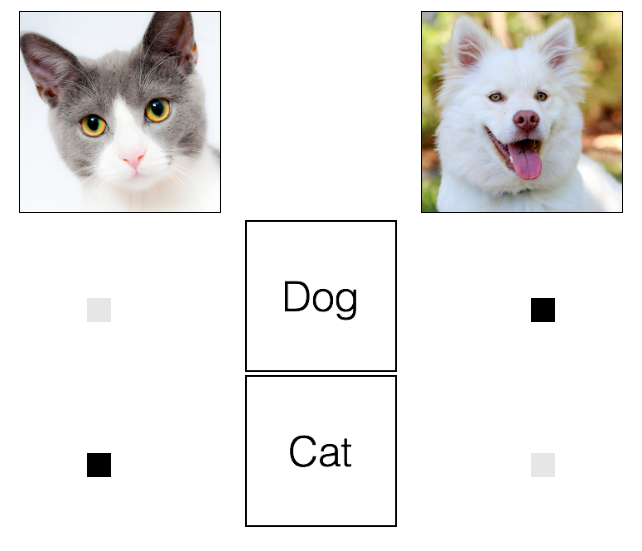
For representations of emotions, we used emoji-like figures taken from a variety of sources via google image search and marked as free to use, share, or modify. The sets of images were then standardized for size and converted to greyscale with equal brightness in an attempt to eliminate any effects of color.

### Procedure

#### Familiarisation

After being briefed and providing informed consent, participants were given some practice with the Experimental interface to ensure that they understood how it operated. As a first example, they were presented with two pictures: a cat and a dog, and the English words "Cat" and "Dog" and asked to pair the two words using the interface (see Figure 8 below).

##### This is a line break

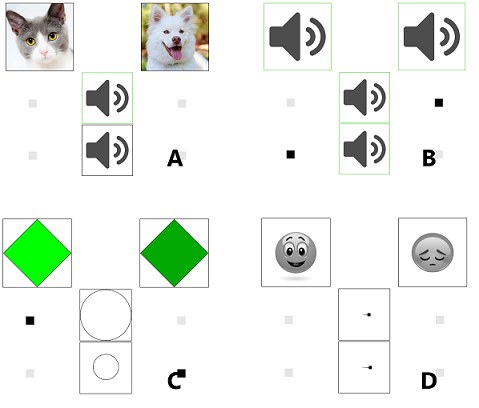


**Figure 8-** *A practice trial familiarising participants with the interface for Experiment 1*

##### This is a line break

After this initial practice, participants were given an additional four familisarisation trials (Figure 9 below) that gave them a preview of the types of visual and auditory stimuli that they would be exposed to in the Experiment. The first two of these familiarisation trials (Figure 9, panels A and B) also served as our first attention checks, as both had sensible correct answers (pairing the sound of a dog with a dog, or pairing two loud sounds together). Participants were informed on those trials that the trial was an attention check, and participants who failed either of those trials were excluded from participating in the Experiment and paid twenty-five cents for their trouble.

##### This is a line break

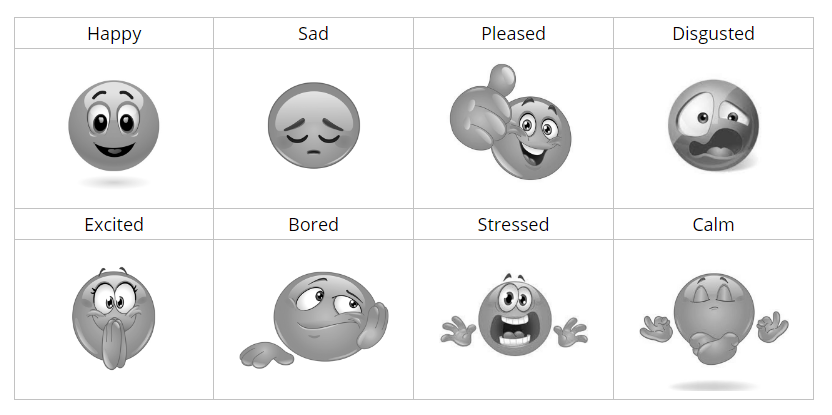


**Figure 9-** *Practice trials from Experiment 1 showing participants a variety of stimuli that were be used in the experiment*

##### This is a line break

Finally, to ensure that participants were familiar with the intended connotations of the emoji expressing emotions, they were shown the full set of emotion stimuli with labels (Figure 10).

##### This is a line break



**Figure 10-** *A familiarisation screen showing participants the set of emotion images and their intended meanings for Experiment 1. In the actual experiment, emotions were shown without labels*

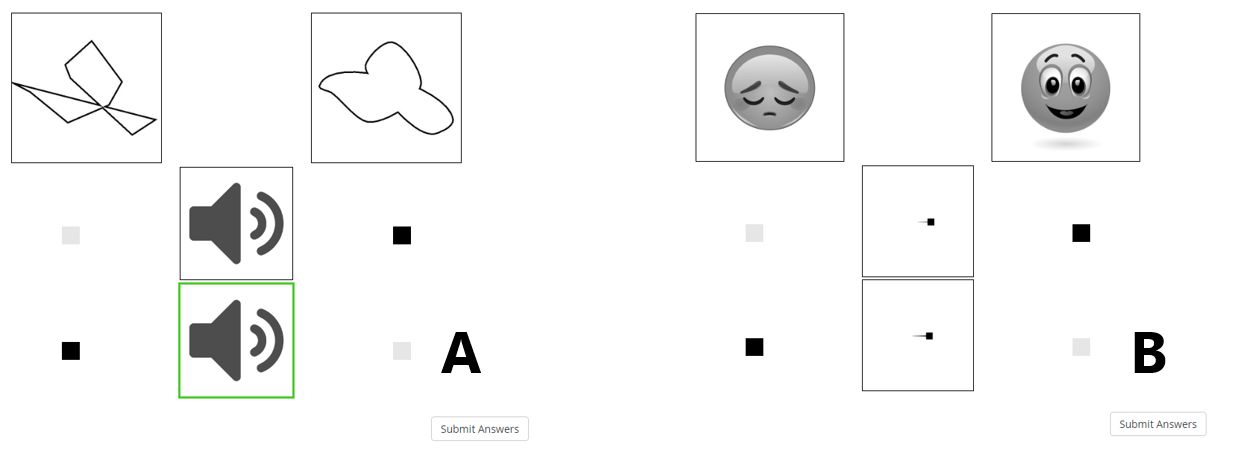
##### This is a line break

#### Testing

Testing trials were identical to familiarisation trials (Figure 11, below): participants were presented with a total of four stimuli and tasked with matching them to one another using on-screen buttons. Crucially, the configuration of the buttons assured that participants could only match one stimuli on the top row with one stimuli in the bottom two. Thus, in Figure 11, panel B it would be impossible for a participant to answer that the fast-moving visual stimuli (top) was both happy (right) and sad (left). Clicking the top right button (as shown in panel B) answered that the fast-moving object was happy, and at the same time automatically matched the slow moving object (bottom) with the sad emoticon (left) (so the same answer could have been produced by clicking the bottom left button, rather than the top right).

On trials with auditory stimuli (e.g. Figure 11, panel A) participants were presented with the stimuli by hovering their mouse over its icon. Once a participant heard the entire audio file, its icon was highlighted in green. To ensure that on audio trials participants listened to all stimuli, they were unable to submit a response until all of the audio stimuli had been listened to in full.

##### This is a line break



**Figure 11-** *Two sample testing trials from Experiment 1*

##### This is a line break

After responding, participants advanced to the next trial by clicking on the "Submit Answers" button in the bottom right corner of the interface.

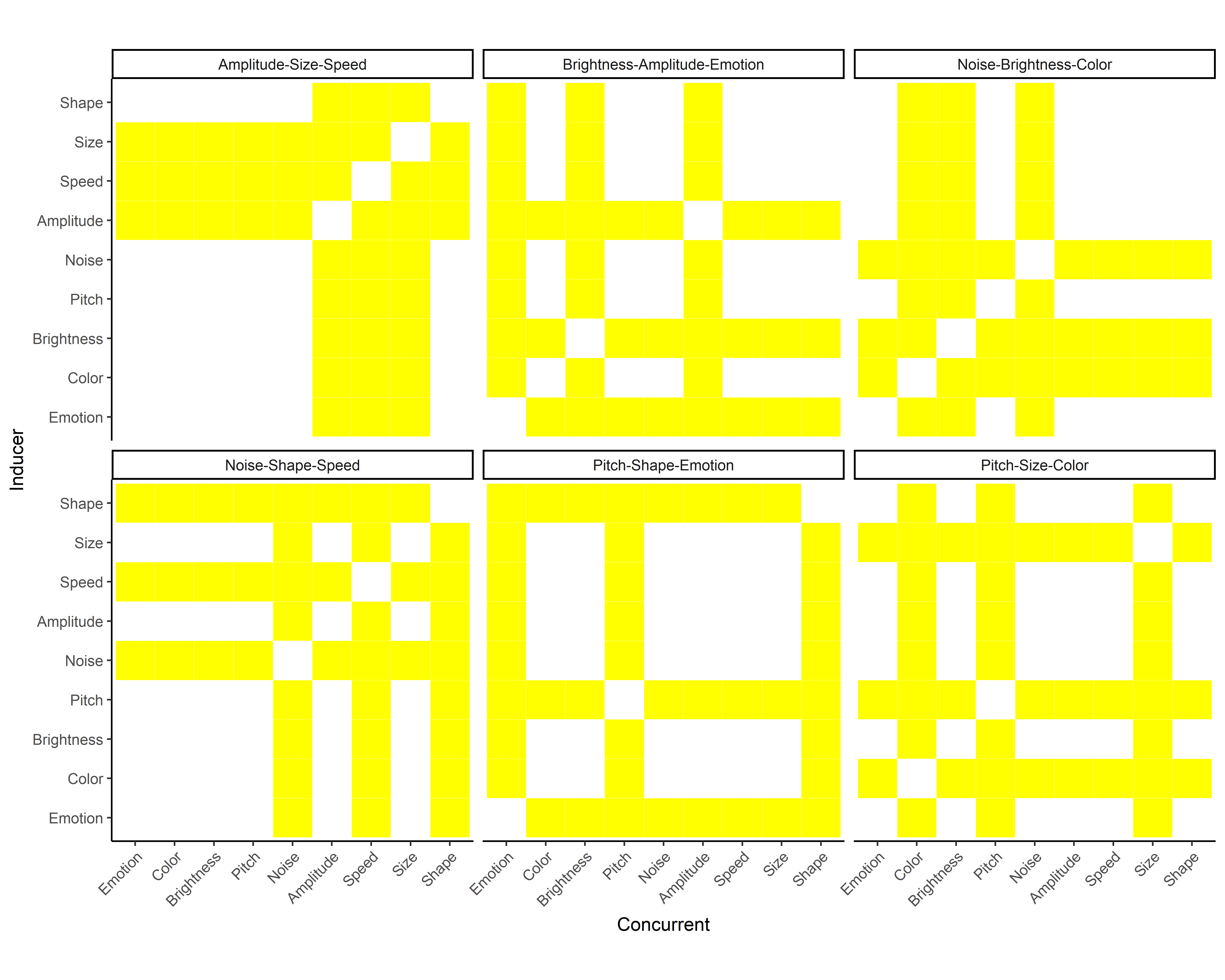
## Experiment Versions

We collected data for several versions of Experiment 1 - an initial pilot to ensure that our stimuli and experimental interface were sensible, a "main" experiment testing associations between all domains (Experiment 1A) and two versions of the experiment focusing on specific domains (Affect in Experiment 1B, Color in Experiment 1C) to bring our total number of trials for each comparison to an acceptable number (and expend our remaining budget). Below, we provide details of the four versions of the experiment and the rationalisation for each. Subsequent to our explanation of the data collection procedure, the main data analysis that follows collapses across the four sources of data.

### Pilot

Given that our experimental interface was novel, and we were also using novel stimuli to make an exhaustive set of comparisons, some of which had not been made before, we chose to pilot our data with an initial group of sixty experimental participants. In this pilot, participants made comparisons between all of the nine domains of interest. Making a complete set of comparisons between all 9 domains however would have required, at minimum 72 trials, which would have given only a single data point on each comparison for each participant. Because we wanted at least some idea of the stability of response from individual participants, and because testing individual participants on all associations in a sufficient number of trials would have resulted in an experiment too long for deployment on mechanical turk, we instead opted to split our participants into six subconditions, each of which was assigned 3 'focal domains' (Figure 12, below).

##### This is a line break



**Figure 12-** *Subconditions of our Pilot and Experiment 1A- for each subcondition the comparisons made for each participant are yellow*

##### This is a line break

For each participant in the pilot, 3 focal domains were compared exhaustively with all other domains, for a total of 24 combinations. This allowed us to test each of those combinations a total of four times per participant, once with each token set for the focal domain. So, for example, a participant with "Shape" as a focal domain would see all four pairs of bouba-kiki stimuli (Figure 6), once each. For non-focal domains, a single token from the set of four was chosen for each participant: for example if "Size" was a non-focal domain, an individual participant might see only Large vs. Small Triangles on all "Size" trials. This resulted in a total of 96 trials for all participants in Experiment 1A.

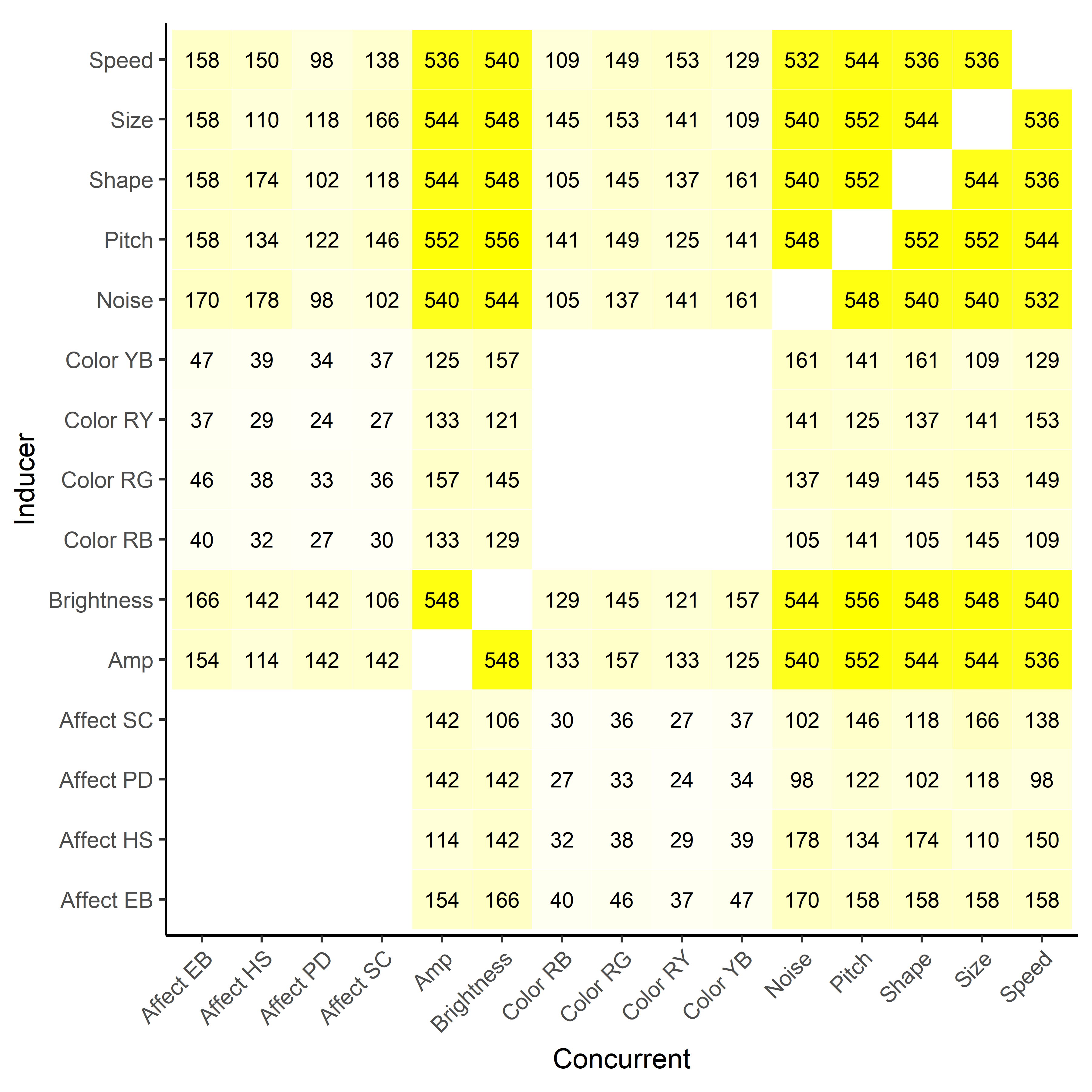
The results of our Pilot were analysed and effect sizes computed (Described in the Data Analysis section below). These computed effect sizes were used to make sample size estimates, which suggested that, on average, we would need a sample size of ~90 for each comparison. Based on a combination of this sample size analysis and a consideration of our approved experimental budget, we split our remaining funds between the three versions of the experiment described below.

### Experiment 1A

Experiment 1A directly mirrors the initial pilot. We collected an additional 144 participants between the 6 experimental subconditions. For analysis, we collapsed the Pilot Data with the Data from Experiment 1A.

For the majority of domains being studied here, we had no explicit predictions about the effects of specific tokens on response biases - for example we don't predict that the four sets of Bouba-kiki stimuli contrasting in shape will be responded to any differently - the choice of four token sets for each domain was thus motivated primarily by a desire to ensure that results were tokens as either Size-High (Large) or Size-Low (Small). This is not true, however, for the domains of Color and Emotion, for which each set of tokens is an entirely different comparison - we may want to make predictions that Happy and Excited are similar based on their valence, but overall it makes much more sense to consider each pair of emotions (e.g. happy vs. sad) alone.

##### This is a line break



**Figure 13-** *Counts of Trial Numbers for Experiment 1A when the Color and Affect domains are broken up by token set*

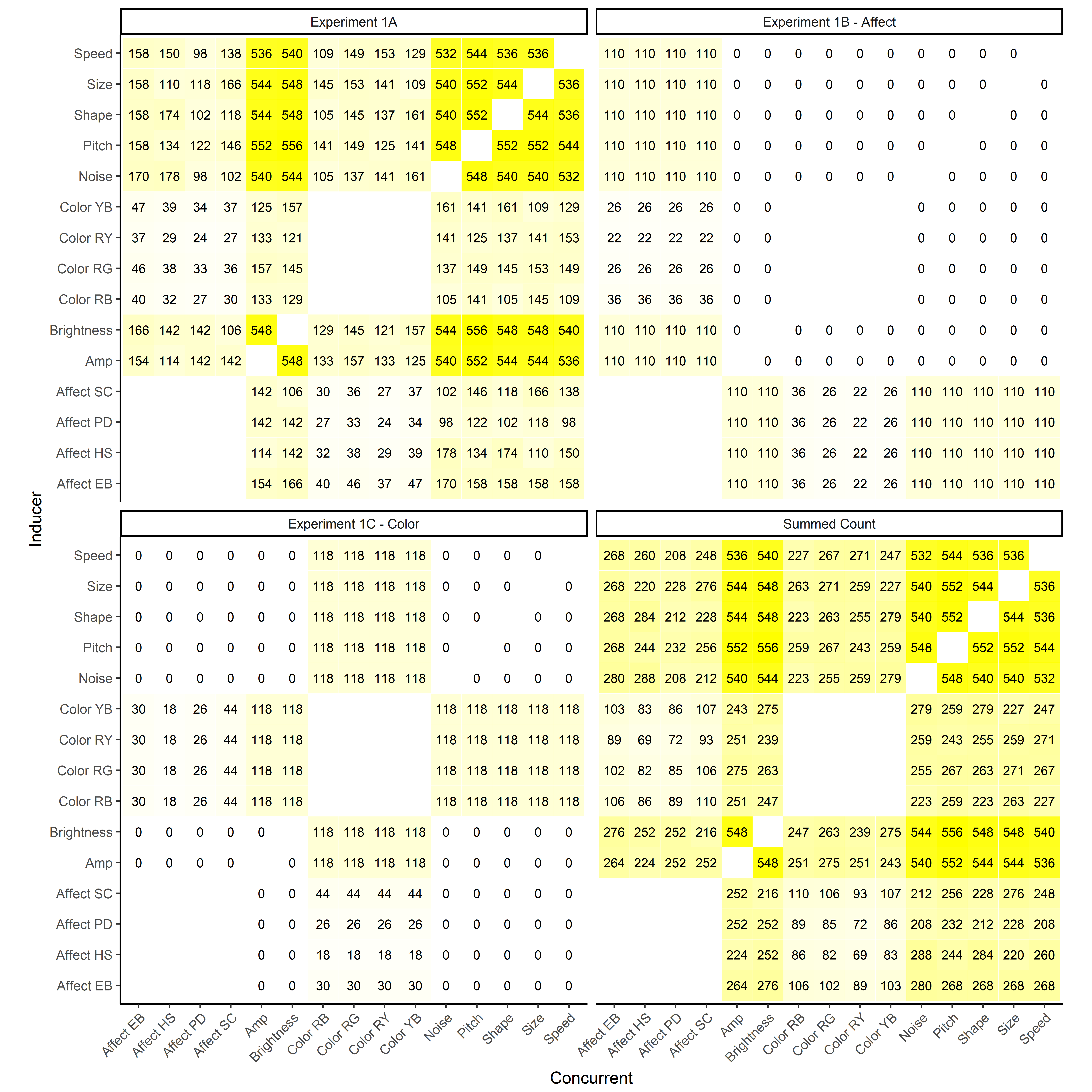
##### This is a line break

As can be seen above in Figure 13, breaking apart Color and Affect into their component tokens means that some comparisons have very few comparisons- as low as 24 total trials. Given that our pilot data suggested that we should have upwards of 100 trials for each comparison, we decided to collect additional sets of data focusing on Affect (Experiment 1B) and Color (Experiment 1C), whichwe describe briefly below.

### Experiment 1B (Affect Focal)

We collected data for an additional 55 participants in a version of Experiment 1 that used Affect as its only focal domain. These comparisons each made two comparisons between all four Affect pairs and all other domains, for a total of 64 trials. Figure 14, below, shows the number of comparisons made for Experiment 1B (Panel B). Otherwise, Experiment 1B was identical to the main experiment.

##### This is a line break



**Figure 14-** *Trial Counts for the three experiments, as well as the summed number of trials for each comparison*

##### This is a line break

### Experiment 1C (Color Focal)

We collected data for an additional 59 participants in a version of Experiment 1 that used only Color as a focal domain. The design of this Color Focal version of the experiment was identical to the design of Experiment 1B, see Figure 14.

## Results

### Data Preparation

I don't know what needs to go here

### Data Analysis

SHOULD THIS BE A SEPARATE SECTION?

#### Sample Size Analysis

For each comparison in our Pilot Data, we used the raw response data (what a participant chose to match with what on a given trial) to compute aggregated responses for every comparison of domains. For example, we found that of 49 trials where participants had to match an Excited face and a Bored face to visual images that different in brightness, they chose to match the Excited face to the Brighter image ~94 % of the time.

For each comparison, we conducted a non-parametric wilcoxon signed ranks test on the raw binary response data, which yielded an alpha value for all 93 comparisons. From this p-value we were able to recover the z test statistic, which we then used to calculate the effect size (Cohen's D) for that comparison following the procedure described by Pallant (2007):

With these effect sizes we used the pwr.t.test function from the pwr package, version 1.2 (Champley et al, 2017) to calculate required sample sizes with 90% power at an alpha level of 0.00054 (p of 0.05 corrected for 93 total comparisons). We estimated required sample sizes for all comparisons that were significant at an alpha of 0.05 in the pilot dataset: 56 out of 93 total comparisons. This conservative estimate suggested a range of required sample sizes, from a minimum of 28 for the comparisons of Affect (Happy vs. Sad) to Brightness (to which 100% of responses suggested that Happy = Bright) to a maximum of 529 for the comparison of Noise - Pitch, to which ~60% of responses suggested that High Pitch = Noisy.

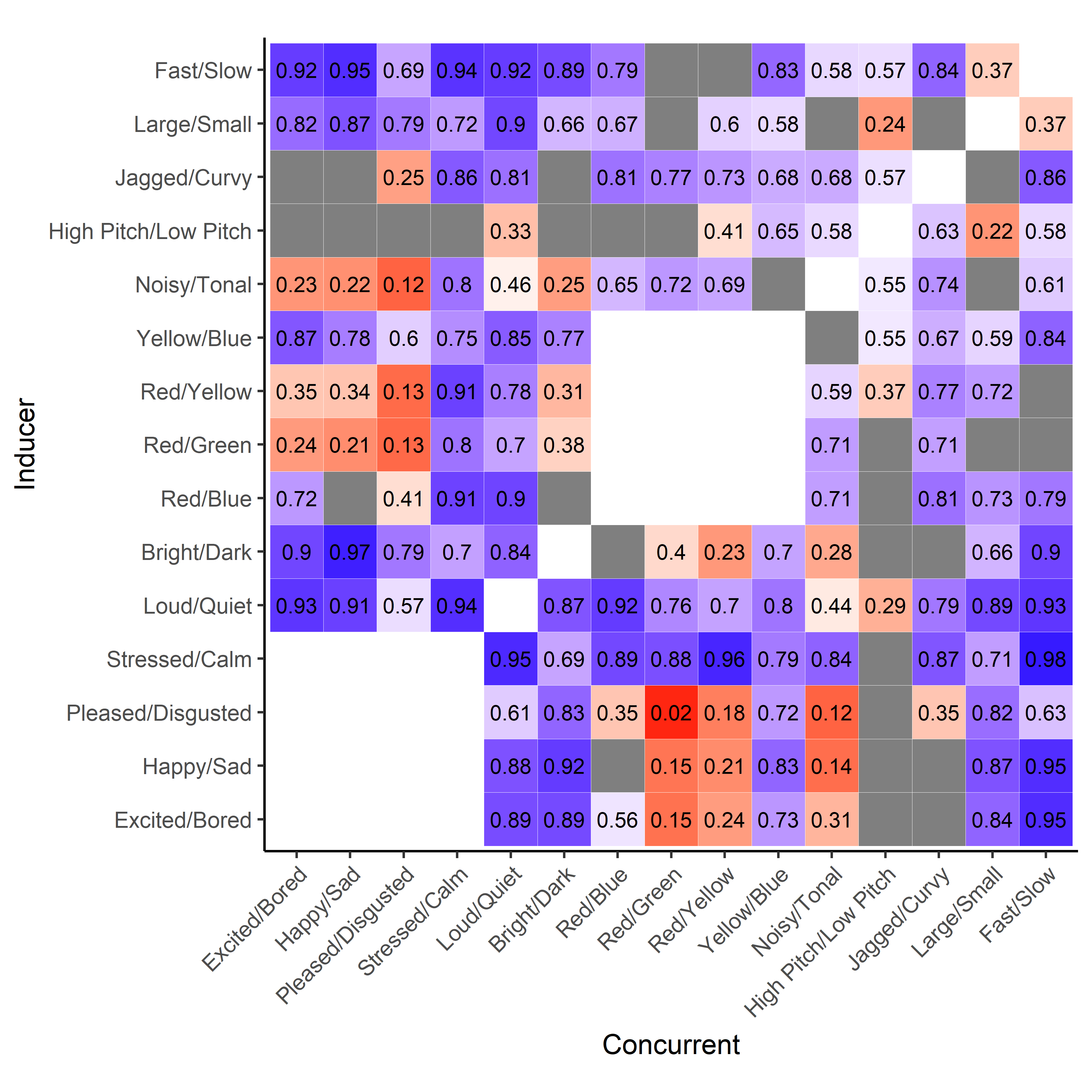
Given this range, we opted to use the mean required sample size as our minimum number of observations for comparisons - this suggested that we should have at least 97 observations per cell, a condition that we met for almost all comparisons, save some comparisons between emotions and specific colors.

#### Descriptive Statistics

The majority of our interest in this data set was purely descriptive - we wanted to catalogue associations between a number of domains exhaustively and using a shared set of stimuli. To that end, we collapsed data from all four sources (Pilot, Experiments 1A - 1C) for analysis. Data analysis followd the procedure described above for the Pilot data- for each comparison we used a Wilcoxon-signed ranks test on the raw response data, giving us an overall alpha value for each of the 93 total comparisons.

In order to correct for multiple comparisons, we emply the Benjamini-Hochberg method (Ref). After applying this correction, we found that 2.5 of 93 comparisons gave results that were significantly different from chance (without correcting for multiple comparisons, 2.5 of comparisons were significant, while the stricter Bonferonni correction method suggested that `r length(BonSig)/2 comparisons yielded significant results). The overall results can be seen below in Figure 15.

##### This is a line break



**Figure 15-** *Heatmap of overall results from Experiment 1. Color and text indicate the proportion of responses that matched the Inducer (Top set of figures in experimental paradigm) listed on the left to the concurrent listed on the left. Thus, for example, the bright blue square in the top corner indicates that on 92% of trials participants suggested that Fast = Excited (and Slow = Bored). In the Pleased/Disgusted row you can see an example of a significantly negative association- participants almost never matched Pleased with Red (2% of trials), instead suggesting that Pleased = Green (and Disgusted = Red).*

##### This is a line break

CURRENTLY WE AREN'T REPORTING EFFECT SIZES- IN THE PAST I HAVE PLOTTED THESE IN ADDITION TO THE PROPORTION OF RESPONSES- JUSTIN MAYBE YOU'D LIKE TO HAVE A THINK ABOUT HOW BEST TO PLOT THIS?

#### Other Statistics

I DONT' KNOW WHAT WE WANT TO INCLUDE HERE- I SUGGEST NOT MUCH, IF ANYTHING AT ALL. I KNOW WE HAVE DISCUSSED PUTTING IN SOME OF THE FACTOR ANALYSIS ETC, BUT NOW I'M THINKING MAYBE NOT, ESPECIALLY IF WE WANT TO HAVE A HOPE IN HELL OF MAKING A 3000 WORD LIMIT FOR BIG JOURNALS

## Discussion

The results presented here are the first attempt at a large scale test of crossmodal correspondences using a shared set of stimuli and the same experimental participants. Many of the results that we find are in line with previously reported correspondences - like the well known correspondence between size and pitch (Small = High Pitched), while others are, to the best of our knowledge, entirely novel (e.g. the observed association between Brightness and Speed).

Perhaps the most surprising of our results is the fact that we observe so few associations between pitch and other domains, despite the fact that pitch is one of the most broadly studied domains for crossmodal correspondence. The overall lack of pitch associations may verify that, indeed, there has been some bias in the choice of associations tested in the past, but may also demonstrate that our stimuli are imperfect, which we recognize as a possibility.

Overall, we suggest that the approach presented here is a valuable one, as having a network of assocations recorded using the same methodology should allow us to compare results and make sensible predictions about how different crossmodal correspondences are related to each other. Now, rather than attempting to piece together explanations from individual experiments that might differ in their stimuli, experimental methodology, and participants (as in Figure 2), we can make direct comparisons using this newly collected dataset, and whatever future, improved versions that the crossmodal research community is able to collect.

Below we discuss a number of future directions for this type of systematic approach to studying crossmodal correspondences.

### Future Directions

#### Stimuli

The stimuli for this experiment were chosen primarily based on how easy there were to create and deploy in an online settings. In some cases, like our stimuli that differed in "color", this meant simplifying a complex psychophysical dimension into a binary division, while in others like "emotion" this meant using conventionalised emoji representations of emotions, which are somewhat far removed from both a) actual emotional states and b) facial expressions of their associated emotions. Finally, in all cases our stimuli were binary and simple - a pair of images could differ along only a single dimension, rather than being more like natural objects that can be any combination of size, shape, color, etc. Expanding the exploration of crossmodal associations, both those we included here and in novel stimulus domains, is a natural extension of this work.

#### Task Differences

The forced choice task used here is fairly standard in the psycholinguistic literature, where these sorts of tasks are commonly employed when looking at size-sound symbolism or the bouba-kiki effect, but one that is less popular in the crossmodal correspondence literature in general. There are a number of reasons for this, the main one being that by virtue of being so explicit, we might measure task demand effects rather than real biases in how human participants associate stimuli from across domains. One benefit of our approach is that it is rather simple to deploy online, allowing for the collection of large amounts of data, but we recognize that extending work with these stimuli to different tasks is a pressing concern.

It may ultimately be the case that implicit tasks are more appropriate, but, echnoing Westbury (2005) we suggest that performance differences between implicit and explicit tasks might inform us about different types of crossmodal correspondences, rather than treating all crossmodal correspondences as being isolated and unitary phenomena. From this perspective, the results of our explicit task can serve as a baseline for comparison to subsequently collected data using more implicit procedures.

#### Between Group Differences - Developmental, Crosslinguistic, and Crosscultural

Our experimental results were collected with adult participants who are not language naive and are generally (by dint of being Mechanical Turk workers) experienced with participating in experiments. Many researchers have speculated, for a variety of reasons, that we should not expect that adult participants would respond similarly to younger participants (e.g. Maurer, Pathman, & Mondloch, 2006). Whether this is true because there are developmental differences in sensitivity to crossmodal associations (ref) or simply because adults, by virtue of experience, are more sensitive to task demands and responding systematically is currently not particularly well known.

Similarly, because our experiment was putatively conducted using English-speaking WEIRD (ref) participants, it is difficult to determine whether the crossmodal correspondences that we have uncovered are truly universals of human perception, or whether they are some way contingent on language and/or other cultural practices. For example, many of the strongest associations found in our dataset are those between colors and emotions: Happy is Yellow, Sad is Blue, Stressed is Red, Pleased is Green, etc. Many color-emotion associations are encoded in English idioms like "feeling blue", "seeing red", "green with envy", etc. (Steinvall, 2007), and thus it might be unsurprising when they crop up in the results of these sorts of experiments. This does not necessarily mean, however, that the source of these observed crossmodal correspondences is linguistic - it might be the case that even in languages that do not have linguistic color-emotion associations the same correspondences are still observed in the language's speakers.

#### Imputation and Toolkit Building

THIS IS ANOTHER SECTION WHERE I'M NOT SURE HOW MUCH WE WANT TO INCLUDE- I'M IN FAVOR OF MAKING THE IMPUTATION STUFF A TOTALLY SEPARATE PAPER, HONESTLY, BUT MAYBE WE NEED PROOF OF CONCEPT HERE?

One additional strength of the approach here is the ability of this type of large network of associations to make predictions based on new data. We might, for example, take data from existing experiments showing that participants associate high-front vowels with small jagged images (Westbury et al., 2017) and feed this information to our network to make predictions about how vowel height and frontness might be associated with other domains like brightness, speed, color, or emotion (cf. Thompson, Nielsen, & Sulik, in prep). The ability to impute these kinds of predictions allows this kind of approach to very quickly build upon itself, producing an increasingly rich dataset that will allow for more advanced techniques attempting to explore the relationships between crossmodal correspondences (Sulik, Thompson, & Nielsen, in prep).

### Conclusions

Much has been made of crossmodal correspondences as potential universals of human perception and cognition. The findings of research conducted over the past century has made this an attractive proposition: as has been noted numerous times, it appears that wherever we look for crossmodal correspondences we find them. The simple cataloguing of individual crossmodal correspondences using disparate sets of stimuli, participants, and experimental designs however obscures our ability to answer the truly interesting questions about crossmodal correspondences:

1- What crossmodal correspondences do we have (or perhaps more surprisingly, what crossmodal correspondences *don't* we have?)? 2- To what degree are crossmodal correspondences shared? a- between members of the same population b- between children and adults c- between speakers of different languages d- between people from different cultures 3- How are crossmodal correspondences related to each other? 4- How are crossmodal correspondences in the general population related to synesthesia? 5- What influence do crossmodal correspondences have on language and other cultural practices?

All of these are important questions with broad implications for the cognitive sciences, but the practice of stamp collecting in the crossmodal literature makes them difficult if not impossible to answer. Even the rise of metaanalysis and other techniques for dealing with and comparing data collected via disparate methods cannot fully ameliorate this problem. Here, we suggest that the best approach is the standardize stimuli and methods, making data between studies more comparable in an attempt to build up a network of associations - a crossmodality toolkit - that can be used to attempt to tackle these questions.

The data presented here are a modest first attempt at doing so, and are, naturally, imperfect. Stimuli need to be expanded and strengthened, task demands need to be considered, additional domains need to be explored and incorporated, etc. However, we suggest that even for the most skeptical of readers the approach we have taken here is a valuable one, because the richness of this dataset makes comparisons so straightforward. In the past attempting to explain why a single study found the results in did could be impossible, with so many factors conflated, but comparison to our dataset using, for example, the same stimuli but a different task (or the same task but different stimuli) should allow researchers a better shot at isolating the effects of individual changes.

AND SOMETHING A BIT PITHIER GOES HERE

##### This is a Page Break

# Acknowledgements:

The design and completion of the experiments presented here were part of a large collaboration between the Language Evolution and Interaction Scholars of Nijmegen at the Max Planck Institute for Psycholinguistics. Experiments 1A and 1B were conducted as part of the September Tutorial in Empiricism, a summer school for early-career academics and part of the Minds, Mechanisms, and Interaction in the Evolution of Language Workshop. Data were collected online during the course of the summer school and analysed live with the summer school participants, who we would like to thank for their patience and participation.

##### This is a Page Break

# References: