Connecting Crossmodal Perception and Garner's Integrality: An Investigation into Crossmodal Correspondences using General Recognition Theory.

Steven Keith Scheid
Chantilly, VA

Master of Arts, University of Virginia, 2014
Bachelor of Arts, George Mason University, 2011

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Psychology

University of Virginia
December 2017

Contents

Introduction

A crossmodal (CM) correspondence is an enduring association between two stimulus features in different sense modalities in which the presence of a stimulus from one dimension can affect the way an individual perceives or responds to a stimulus from the other (Spence & Parise, 2012). For example, in matching experiments, participants are more likely to match brighter visual stimuli with louder auditory stimuli, and speeded classification tasks have found a consistent RT advantage for compound stimuli that match brighter and louder components together (Marks, 1987, 1989).

Speeded classification tasks, where participants are instructed to classify a random compound stimulus based on only one of its two components, are one of the more popular ways to test CM correspondences (e.g., Evans & Treisman, 2010; Gallace & Spence, 2006; Gebuis & van der Smagt, 2011; Marks, 1987). In a brightness–loudness experiment, participants would be shown a compound stimulus made up of a randomly selected pair of visual and auditory components, and instructed to classify the stimulus based on only one of the two stimulus dimensions. If a correspondence exists, participants are faster at sorting stimuli based on a single dimension on trials where both stimulus components match.

However, while speeded classification tasks are useful in identifying which correspondences exist, they do not provide much insight into how CM correspondences operate. Because speeded classification tasks provide a limited amount of data, they cannot answer the question of whether changes in RT stem from genuine crossmodal influences on the perceptual process, a change in response strategies, or both. While the majority of theoretical work on CM correspondences assume that they influence processes related to the perception of the stimulus components, rather than affecting the individuals' decision strategies (De Gelder & Bertelson, 2003; Spence, 2011), the literature is far from unanimous (Marks, Ben-Artzi, & Lakatos, 2003). The lack of data on this critical point impedes any attempt to offer a comprehensive theory about what correspondences are and why they exist.

An alternative is General Recognition Theory (GRT, Ashby & Townsend, 1986), a multivariate extension of Signal Detection Theory (SDT, Green & Swets, 1966; Macmillan & Creelman, 2004), which provides the theoretical framework to separately model both perceptual- and decision-level effects. The GRT model describes a 2×2 classification task as a two-stage process, describing both the extent to which the stimulus dimensions interact as they are being perceived, as well as the decision rules that determine which response is chosen. By separately modeling perceptual and response-level effects in the analysis, GRT can directly address the question of at what stage crossmodal effects take place, as well as providing a description of how these effects influence the responses made by participants.

GRT defines two types of perceptual-level interactions and one decision-level interaction which can be identified with a standard 2×2 GRT experiment. Perceptual integrality, as defined by GRT, describes a situation where two stimulus dimensions interact in such a way that the mean perceived value of a stimulus dimension, or the variability of this value over repeated presentations, changes as a function of the other stimulus dimension (Ashby & Townsend, 1986; Soto, Vucovich, Musgrave, & Ashby, 2015). Perceptual independence describes the relationship between stimulus noise across the two dimensions; like SDT, GRT assumes that perception is inherently noisy, and that there will always be some degree of perceptual noise associated with the process (Ashby & Soto, 2014; Ashby & Townsend, 1986; Green & Swets, 1966). The question of perceptual independence refers to whether the perceptual noise values in each dimension are statistically independent; when perceptual independence does not hold, the two noise components are not statistically independent which reveals an interaction between the stimulus dimensions at the perceptual level (Ashby & Townsend, 1986). The ability to detect if either of these perceptual-level interactions exist for CM correspondences would fill a gap in the CM literature.

At the decision level, GRT models a participant's response strategy as a pair of decision rules, one for each dimension. If decisional separability holds for both stimulus dimensions, then the participant's response for each

dimension is based solely on his or her perception of that dimension's stimulus component, e.g., in a brightness–loudness experiment, the decision between a "bright" or "dim" response is based only on the perceived level of brightness, and not on how loud the auditory component is perceived to be. For example, if decisional separability does not hold for brightness, a participant may be more willing to make a "bright" response when the auditory component is loud rather than soft, or vice versa. It is important to note is that while perceptual separability and independence are considered to be enduring traits, decisional separability reflects the strategy employed for the task at hand, and so there is no expectation that it would generalize to other contexts (De Gelder & Bertelson, 2003; Soto et al., 2015; for those who are interested, a more thorough explanation of the methods and analyses of GRT is given in the Appendix).

GRT experiments also make it possible to measure the discriminability of compound stimuli, which can be used to evaluate the idea that CM correspondences represent an adaptive strategy. Some researchers hypothesize that CM correspondences have ecological value by allowing an organism to improve perceptual performance by taking advantage of multiple redundant sources of information (De Gelder & Bertelson, 2003; Spence, 2011). However, actual tests of CM congruency and discriminability have been scarce (e.g., Chen & Spence, 2010; Sanabria, Spence, & Soto-Faraco, 2007), and the majority of CM studies do not address this topic. As GRT is an extension of SDT (Ashby & Townsend, 1986; Green & Swets, 1966; Macmillan & Creelman, 2004), the sensitivity statistic (d') from SDT can be adapted to measure the discriminability of two compound stimuli. By adopting the GRT experimental design, the relationship between perceptual sensitivity and CM congruency can be evaluated for each correspondence tested, without the need to collect additional data.

The richer dataset generated by GRT experiments also makes it possible to identify what features are common to some or all CM correspondences, and whether attempts to define certain subtypes of correspondences are supported by empirical data. The most well-developed theoretical taxonomy of CM correspondences comes from Spence (2011); in his paper, Spence (2011) groups correspondences into three non-exclusive categories based on both the characteristics of the stimulus dimensions and of the perceptual system, and describes how these characteristics may give rise to various crossmodal correspondences. The first category is Statistical, which describes correspondences that appear to be reflections of naturally occurring statistical regularities from the environment. Correspondences in this category are the most likely to be advantageous to the organism, as they represent a strategy to account for reliable patterns in multimodal stimuli in the external world. The second category is Structural, which describes pairs of stimuli that become associated due to shared neural pathways or other innate connections in the brain. This category would contain the correspondences which appear to be based on magnitude (Spence, 2011; Walsh, 2003). The third category is Semantic, which refers to pairs of stimulus dimensions that are related due to common verbal descriptors or conceptual associations. Correspondences of this type are thought not to be associated with one another at the perceptual level, but still show a congruency effect because their semantic associations can influence response strategies.

While Spence's taxonomy provides a way to organize the wide range of disparate experimental results and methodologies, it has not been subject to a proper empirical test. To some extent, this is likely due to the limitations of the methodologies used in the field: speeded classification experiments do not provide enough data to distinguish between different types of correspondences. The same can be said for the majority of published research, where the goal is to demonstrate that a given correspondence exists, rather than investigate how it behaves. Because the GRT paradigm produces a richer set of data, Spence's theoretical divisions can be compared to patterns that appear in the results; if the results show patterns that match Spence's categories, it would support the classification scheme he has developed.

Brightness–Loudness

The first experiment tests the correspondence between visual brightness and auditory loudness. The link between brighter objects and louder tones in a speeded classification task is well established (Marks, 1987), and appears to be a prototypical magnitude-based correspondence (Gallace & Spence, 2006; Marks, 1987, 1989; Spence, 2011). This makes this correspondence ideal for testing with GRT: the correspondence has strong empirical support, a clear rationale for why this correspondence exists, and the correspondence is unlikely to be caused by language-based associations.

Additionally, this experiment will test whether this correspondence produces an increase in discriminability for congruent compound stimuli. The presence of a correspondence, regardless of type, does not necessarily imply an increase in perceptual sensitivity and unlike the size–loudness or size–pitch correspondences, there is no regularly occurring association between brightness and loudness in the environment, and so there is no ecological reason for a congruency advantage to be present. However, if a relationship between crossmodal congruency and perceptual sensitivity exists, it should appear in the results of the sensitivity analysis.

Methods

All seven experiments follow the same 2×2 classification task template; the first experiment is described in detail and subsequent experiments are presented in an abbreviated format.

Participants. Participants were recruited from a pool of undergraduates at a large research university and participated for course credit. Participants who were unable to complete at least 75% of the total number of trials were dropped from the analysis. Additionally, participants who did not perform with sufficient accuracy (defined as having an overall accuracy rate above 45%) were removed.

Outlier detection was done with the individual differences measures from the GRT-wIND analysis (GRT-with Individual Differences, a type of GRT analysis that estimates several individual difference parameters, including a pair of decision bounds for each participant; Soto et al., 2015). When individual decision bound parameters were estimated for each participant, some participants showed a clear deviation from the rest of the sample (Figure 1 is a plot of the decision bound slopes for the loudness dimension using the unfiltered data). These deviations could be the result of at least one of two factors: either these participants were using an unusual strategy throughout the task or there was insufficient data to correctly estimate the response strategy they used. The criterion for removal was for participants with a decision bound slope estimate that was more than three median absolute deviations (MADs) from the median slope. For all but one experiment, this produced no substantive change in the findings, and the exception is noted in the appropriate section.

Data were collected from 67 participants. Three individuals were removed for not completing enough trials, and eight were removed for insufficient overall task accuracy, leaving a sample of 56 participants. Four participants had decision bound slope estimates more than three MADs from the median and were excluded (one participant was excluded based on the brightness decision bound slope, the other three were removed because of their loudness decision bound slopes), producing a final sample of 52: 21 men, 31 women; median age was 18.

Stimuli. In this experiment, stimulus values were not fixed from trial to trial. Stimulus values for each dimension were generated by sampling from a normal Gaussian distribution with one of two mean values and a set standard deviation, following the methods used in Ashby and Gott (1988) and Ashby and Maddox (1990). This generated stimuli with a range of possible values for each level of the stimulus dimension.

The advantage to intentionally adding jitter to the stimuli is that it allows for the discriminability of the stimuli to be manipulated, as increasing or decreasing the amount of variance in the stimuli would make the task more or less difficult. This method can ensure that all stimulus dimensions are relatively equal in discriminability; if classifying one stimulus dimension is much easier than the other, that could artificially induce an asymmetrical

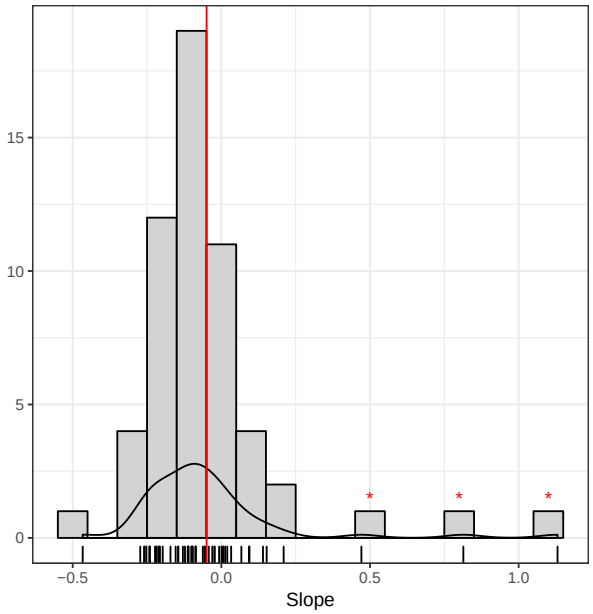

Figure 1. A diagram of the loudness decision bound slope estimates from the brightness–loudness experiment. Three individuals clearly deviate from the rest of the sample (and are labelled with an asterisk).

pattern of integrality, regardless of actual relationship between the two dimensions (Algom & Fitousi, 2016; Garner, 1983). Additionally, the stimulus jitter meant that individual stimuli needed to be evaluated, as opposed to recognized, on every trial, so participants remained focused throughout the experiment.

The parameters of the visual and audio components of the compound stimuli are described in a separate section.

Apparatus. The experiments were carried out on Apple Mac mini computers running Matlab version R2012b (MATLAB, 2012). Each computer was equipped with a 19-inch LCD flatscreen computer monitor and Sennheiser HD 555 headphones. The experimental program was written using Psychtoolbox–3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007; Pelli, 1997).

Task. An experimental session consisted of 12 blocks of 50 trials each (following the example of Soto et al., 2015). On each trial, one of the four stimulus categories was selected and a pair of stimulus values were generated to create a new compound stimulus. All four stimulus categories were equally likely to appear on any given trial, and the particular stimulus values for each component were chosen randomly, so there was no overall correlation between the values on the two stimulus dimensions.

The flow of each trial was adapted from Soto et al. (2015). Each trial began with a fixation cross, which remained on screen until the participant pressed the spacebar. Then the cross was removed and the screen was left blank for 500 ms, after which the stimulus was displayed for a brief interval (16.66 ms, which is the duration of a single frame on a 60 Hz monitor), before the stimulus was removed; the screen then remained blank until a response was made. Stimuli were not masked after presentation; masks are not often used in GRT experiments (e.g., Ashby & Gott, 1988; Ashby & Maddox, 1990; Soto et al., 2015) and were not used here to ensure that the results would be compatible with previous research. The participant then pressed one of four keys (D, F, J, or K) to indicate which of the four types of compound stimuli he or she believed was presented on that trial. After the response was made, participants received feedback indicating if their response was correct or incorrect for 500ms, and the next trial began.

Procedure. The experiment began with an explanation of experimental task, after which each participant was placed in a separate room to begin the experiment. The first screen of the experiment program described which keys were associated with which stimulus categories (as this was randomized between subjects). The experiment was self-paced, and participants typically finished in 45 to 50 minutes.

Results

GRT is intended to model the behavior of expert classifiers (Ashby & Soto, 2014; Ashby & Townsend, 1986; Soto et al., 2015); therefore, before analyzing the data, it is necessary to remove trials where participants are still learning the task and any mistakes that are made may be unrelated to perceptual processes. This was done with a similar method to what was used in Soto et al. (2015): a series of generalized linear models (GLMs) were fit to 100-trial moving windows in each participant's data using correct/incorrect as the outcome variable (creating 501 windows for a 600-trial data set). Then, the slopes of the GLMs were compared to determine where task performance leveled off, which was used to separate the initial learning curve from trials after participants had mastered the task. When 50 consecutive GLMs had a mean slope of less than 0.001, the beginning of that set was designated as the first trial to be included and all subsequent trials were retained to make that individual's filtered data set (the typical number of trials removed with this procedure was around 150, similar to the number from Soto et al., 2015).

Fitting a GRT-wIND model and testing for perceptual separability, perceptual independence, and decisional separability can be done in R with the **grtools** package (R Core Team, 2014; Soto & Zheng, 2016; Soto, Zheng, Fonseca, & Ashby, 2016; a general overview of GRT, as well as the specifics of the GRT-wIND analysis, are covered in greater detail in the Appendix). The model was fit using the **grt_wind_fit_parallel()** function with 200 repetitions (i.e., the model is fit 200 times and the function returns the parameters of the model with the best fit). Once the model was fit, tests of perceptual separability, etc., were carried out via the **lr_test()** function (with a *nreps* value of 50, i.e., repeating the model fitting procedure 50 times to find the best fit). This function fits another set of models with various constraints in place; for example, constraining the first stimulus dimension to show perceptual separability and rerunning the model to determine if this produces a significant reduction in model fit.[1]

Accuracy. After filtering, the median number of completed trials was 507 per participant. Median overall accuracy for all participants was 60.10%. Accuracy for congruent and incongruent stimuli were similar (62.62% and 59.66%, respectively). Looking at accuracy by stimulus dimension, participants made the correct response with regard to the brightness component 80.70% of the time. Participants performed similarly on the loudness dimension, making a correct response with regard to loudness 75.69% of the time.

Perceptual Separability. The analysis (the full model is plotted in Fig. 2) indicated an asymmetrical pattern of integrality. Brightness was perceptually separable from loudness, $\chi^2(4) = 4.13, p = 0.389$, but loudness was integral with brightness, $\chi^2(4) = 12.74, p = 0.013$. The mean values appeared largely stable, e.g., the mean perceptual value for a dim stimulus was roughly the same on the brightness dimension across both levels of loudness; this held true for both stimulus dimensions. The evidence for integrality appears mostly in the variances of the stimulus distributions. The distributions' variances in the brightness dimension appeared largely stable between levels of loudness (as one would expect given that it is perceptually separable), but the variances in the loudness dimension showed a large degree of change across different levels of brightness, indicating that the level of brightness influences the amount of variability in the perception of loudness.

[1] As an additional note, tests of perceptual separability, etc., are carried out one at a time, which has some potential implications for the results (see the Appendix for more details).

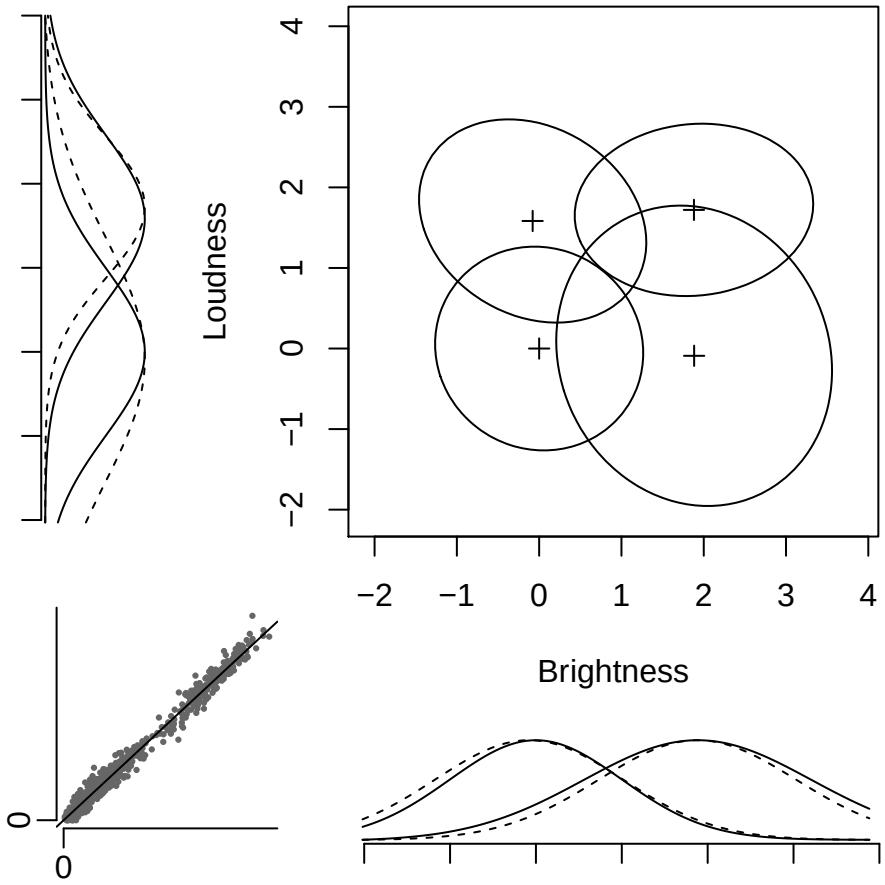

Figure 2. The results of the GRT-wIND analysis for brightness–loudness. For grtools plots, the main plot displays a representation of the participants' common perceptual space; in this experiment, the congruent compound stimuli are represented by the distributions in the bottom left and top right. The figure also includes diagrams of the marginal distributions, and the plot in the bottom left illustrates the relationship between expected and observed probabilities — no overall pattern indicates that there are no trends in the data not captured by the model.

Perceptual Independence. Table 1 shows the correlations between the perceptual values for brightness and loudness for each compound stimulus. Results indicated that perceptual independence does not hold overall, $\chi^2(4) = 29.41, p < 0.001$. The correlations for incongruent stimuli (top left and bottom right) were both negative, and greater in magnitude than those for the congruent stimuli. The negative correlations suggest that for an incongruent compound stimulus, the more the visual element is perceived as being relatively brighter, the more likely the auditory component is to be perceived as relatively softer, suggesting that the perceptual noise present for incongruent stimuli was itself at odds with the direction of the CM correspondence.

Decisional Separability. Tests of decisional separability indicated that it does not hold for either brightness, $\chi^2(52) = 159.94, p < 0.001$, or loudness, $\chi^2(52) = 172.37, p < 0.001$. Participants did not uniformly use unbiased decision rules for either dimension: their criteria for selecting a response for the brightness component of the stimulus was influenced by both brightness and loudness information, and the same was true for selecting a loudness response.

Table 1
Correlations Between Stimulus Dimensions for Each Compound Stimulus

	Dim	Bright
Loud	−0.21	0.07
Soft	−0.01	−0.10

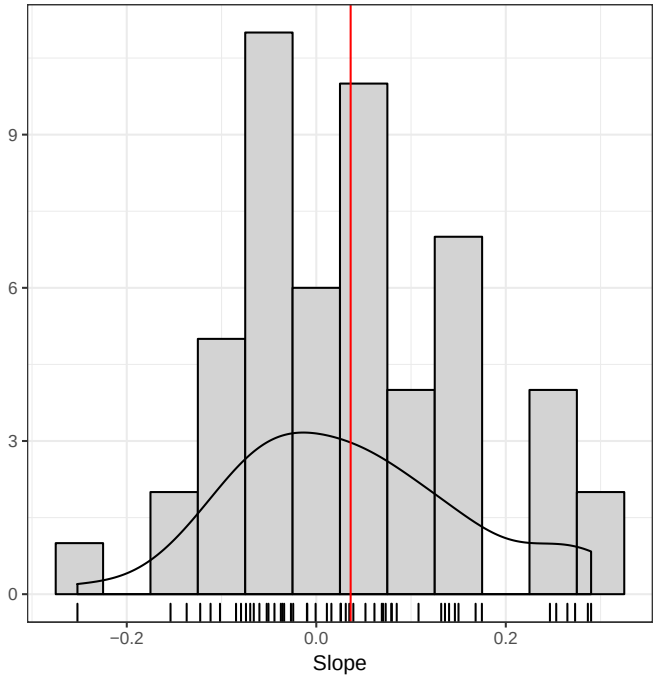

Figure 3. The distribution of slope estimates for the brightness decision bounds. The mean slope value is indicated by a red line.

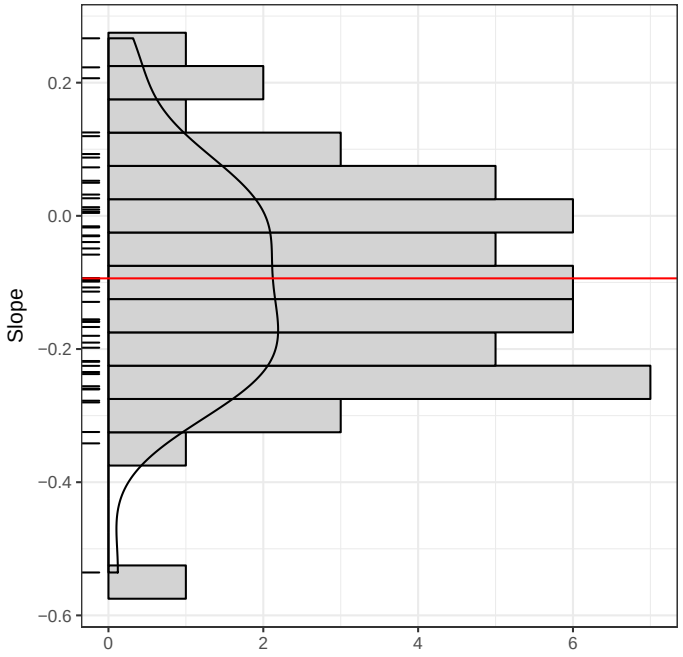

Figure 4. The distribution of slope estimates for the loudness decision bounds. The mean slope value is indicated by a red line.

The participants' collective brightness decision bounds included slopes that were both positive and negative (the values are shown in Fig. 3), but the mean slope value was statistically significantly greater than zero, $\overline{m} = 0.04$, 95% CI: [0.002, 0.07]. This positive slope indicates a decisional bias against classifying something as one of the two congruent stimuli, as the louder the tone is perceived to be, the less likely the participant is to classify the visual component as bright.

The decision bounds for loudness (shown in Fig. 4) likewise included both positive and negative slopes, but had a negative mean slope, $\overline{m} = -0.09$, 95% CI: [−0.14, −0.05]. The negative mean slope indicates that the majority of participants were biased in favor of choosing a congruent response when selecting a response for the loudness component, as brighter perceived values of the visual element increased participants' likelihood of responding "loud."

Sensitivity Measures. Perceptual sensitivity was measured with a bivariate adaptation of SDT's d' statistic. d' is a measure of the distance between two stimulus distributions scaled by the distributions' variance; here, it is defined as the mean Mahalanobis distance between two perceptual distributions (a detailed description of how this was calculated is in the Appendix).

The mean congruency advantage (i.e., the ratio of congruent d' over incongruent d') for all participants was $d'_{ratio} = 1.23$. A bootstrap simulation with 200 repetitions was carried out to determine if the congruency advantage was significantly greater than 1.0; the results gave a 95% confidence interval of [1.07, 1.31], indicating that participants were significantly more sensitive to differences between congruent stimuli compared to incongruent stimuli (the distribution of bootstrapped d'_ratio values is shown in Fig. 5).

This analysis was repeated for the brightness and loudness components of the compound stimuli. The mean congruency advantage for the brightness component was $d'_{ratio} = 1.08$, 95% CI: [0.76, 1.29], indicating that participants were not more sensitive to differences in brightness when the auditory component was congruent. For the loudness component, the mean congruency advantage was $d'_{ratio} = 1.28$, 95% CI: [0.97, 1.70], which also indicates that there was no advantage for the loudness component when it appeared alongside a congruent brightness component.

Discussion

The key finding from this study is that there is an association between brightness and loudness that occurs at the perceptual level, before decision-level processes take place. The findings describe a pattern of integrality between brightness and loudness, albeit an asymmetrical pattern where only the loudness dimension is integral, which demonstrates an early interaction between the two dimensions. Additionally, the analysis of perceptual independence demonstrates an early effect of stimulus congruence, as incongruent compound stimuli showed moderately sized negative correlations between levels of perceptual noise — a pattern which is also found in all subsequent experiments.

The analysis of perceptual sensitivity shows that crossmodal congruency does produce an improvement in participants' ability to distinguish between compound stimuli that differ on both stimulus dimensions, though this advantage only exists for the compound stimuli as a whole and not their components. These results generally support claims the CM correspondences have some degree of ecological utility (Chen & Spence, 2010; De Gelder & Bertelson, 2003; Spence, 2011), though some of the experiments discussed later in this paper have differing results.

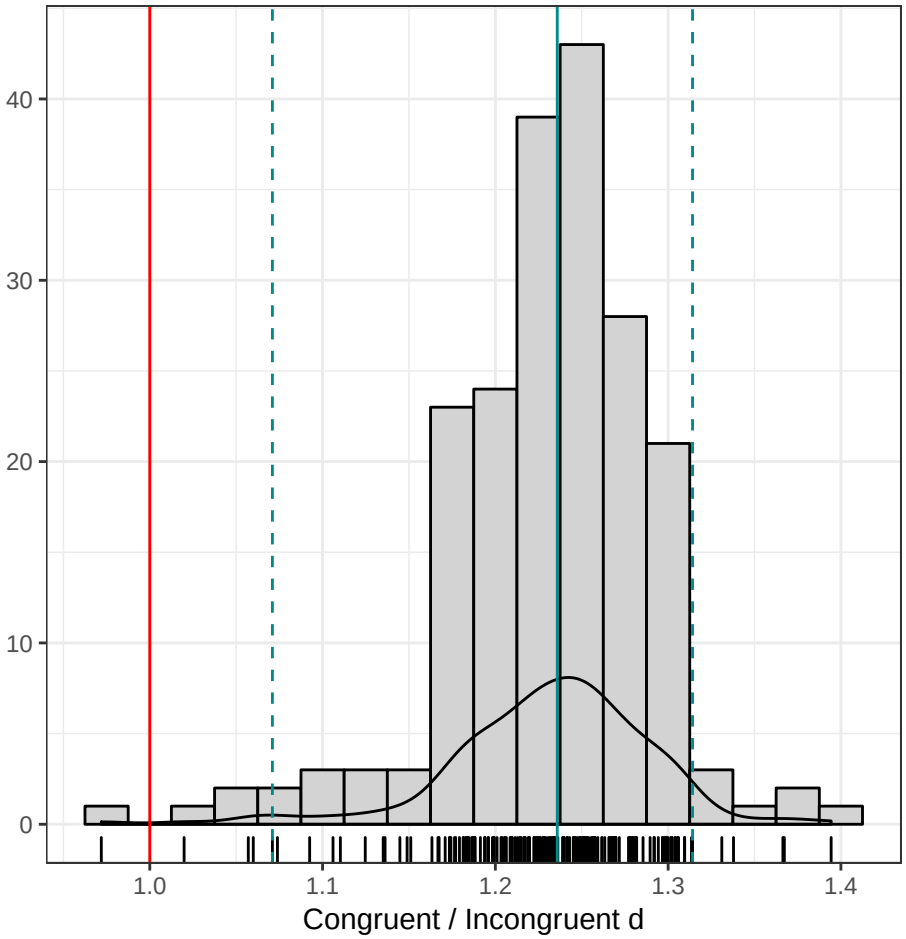

Figure 5. A diagram showing the results of the bootstrap simulation. The solid blue line shows the median congruency advantage for all simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Brightness–Pitch

This experiment examines the correspondence between brightness and auditory pitch, which links brighter stimuli and higher pitched tones (Marks, 1987, 1989). Brightness–pitch is representative of a second type of crossmodal correspondence, one in which the association between the two dimensions cannot be expressed in terms of linking one form of magnitude to another. Brightness and loudness are both prothetic dimensions, meaning changes in stimulus value can be understood as a change in degree (i.e., that stimuli can be thought of as having more brightness or more loudness), but this puts them at odds with metathetic dimensions such as pitch, where different values are perceived as a different type rather than a different amount (Spence, 2011; Stevens, 1957).

The results will be valuable as a comparison to the results of the brightness–loudness experiment. Both correspondences have been thoroughly tested in the speeded classification paradigm, where they produced a similar set of results (Marks, 1987). However, given the limited amount of data generated by the speeded classification task, it is unclear if this similarity was real or an artifact of the task itself.

Methods

Participants. Data were collected from 57 individuals. Five participants were removed for not completing a sufficient number of trials, as were another eleven for insufficient accuracy. One additional participant was removed for having decision bound slope estimates more than three MADs away from the median, resulting in a final sample of 40 participants: 18 men, 22 women; median age was 19.

Results and Discussion

After filtering, the median number of completed trials was 505 per participant. Median overall accuracy was 63.93%; accuracy was similar for congruent (68.82%) and incongruent (62.23%) trials. Accuracy was similar for both brightness (79.77%) and pitch (80.97%).

The results of the GRT-wIND analysis (the model is displayed in Fig. 6) indicated complete perceptual separability. The test of perceptual independence indicated that it did not hold the model as a whole, and, as with brightness–loudness, the incongruent stimuli showed moderate negative correlations. Decisional separability did not hold for either dimension, but the mean decision bound slope was not significantly different from zero in either dimension (the results of this analysis, alongside the results from all other experiments in this paper, can be found in Tables 2, 3, and 4).

The analysis of the sensitivity measures indicate that there was a significant congruency advantage for congruent compound stimuli, $d'_{ratio} = 1.21$, 95% CI: [1.04, 1.31]; the results of the bootstrap analysis are shown in Figure 7. However, neither component of the compound stimuli showed a congruency advantage when analyzed separately (results of the sensitivity analyses are shown in Table 5 for all seven experiments).

In contrast to the previous experiment, brightness–pitch does not show any perceptual integrality, indicating that perceptual integrality is not an essential element of a CM correspondence. However, both stimulus pairs demonstrate a congruency advantage, which indicates that congruency-related advantages in perceptual sensitivity are not limited to magnitude-based correspondences. Additionally, both brightness–loudness and brightness–pitch show the same pattern of perceptual dependence, where incongruent stimuli show moderately sized negative correlations in stimulus noise (the correlation coefficients for all four compound stimuli are listed in Table 3).

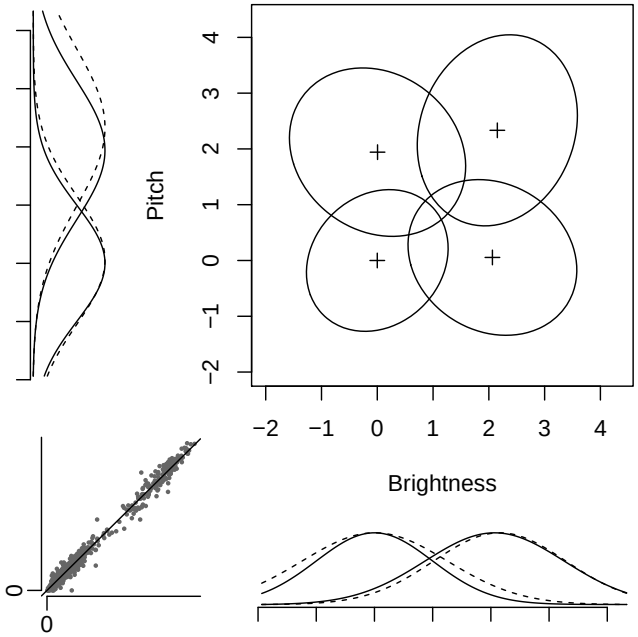

Figure 6. The results of the GRT-wIND analysis for brightness–pitch. The main figure represents the common perceptual distributions of the four compound stimuli (the congruent stimuli are located in the top right and bottom left), displayed alongside the marginal distributions for each stimulus dimension.

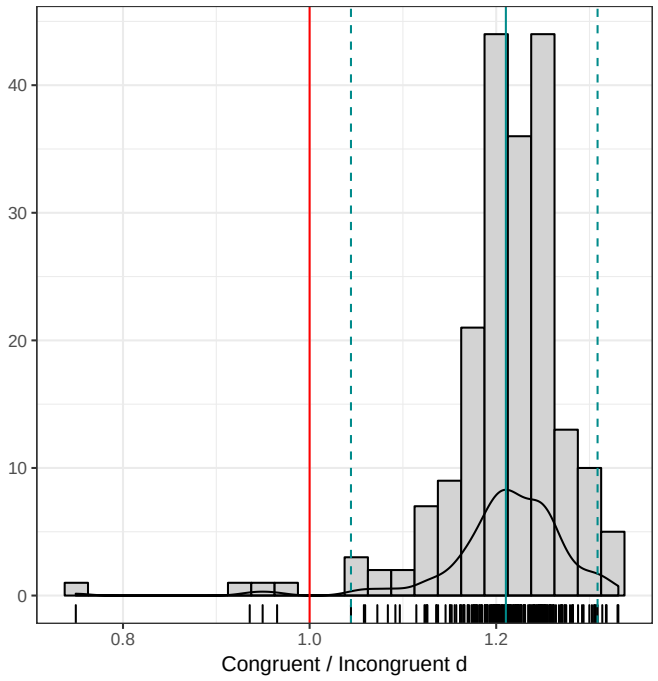

Figure 7. A diagram showing the results of the bootstrap simulation. The solid blue line shows the median congruency advantage for all simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Size–Loudness

This experiment tested the correspondence between visual size and auditory loudness (Walker, 1985), where larger objects are associated with louder sounds. This pairing neatly matches with the idea of shared magnitude coding being responsible for certain crossmodal associations (Spence, 2011; Walsh, 2003), and this explanation gets further support from imaging studies which have identified common brain regions responsible for processing both dimensions (Belin et al., 2002). The results of this experiment provide another look at how magnitude-based correspondences operate.

Methods

Participants. Data were collected from 62 participants (one of which was removed for not completing a sufficient number of trials, and nine more participants were eliminated for not performing with sufficient accuracy). After the analysis was completed, two participants appeared as outliers, and were removed before the final analysis. The final sample size was 50; 18 men, 32 women; median age was 18.

Results and Discussion

After filtering, the median number of completed trials was 545 per participant. Median overall accuracy was 66.50%. Median accuracy for both congruent and incongruent trials was similar (congruent accuracy = 65.02%, incongruent accuracy = 68.32%). Looking at accuracy by stimulus modality, participants' median accuracy for size was 85.20% and median accuracy for loudness was 76.92%.

The GRT model (shown in Fig. 8) demonstrated complete perceptual separability. The analysis of perceptual independence revealed the same pattern in the incongruent stimuli's covariances as was seen in the previous experiments. Additionally, decisional separability did not hold, but the mean slopes estimates for both stimulus dimensions did not differ from zero.

Unlike the previous two experiments, there was no sensitivity advantage for congruent stimuli, $d'_{ratio} = 1.08$, 95% CI: [0.97, 1.18]; the distribution of bootstrapped d'_{ratio} values is shown in Figure 9. Neither component showed a congruency advantage when tested individually.

In contrast to the results of brightness–loudness, there is no evidence of perceptual integrality of any kind, nor is there a congruency advantage in sensitivity. Despite both correspondences appearing to share the same underlying mechanisms (i.e., both are related to magnitude processing in shared brain areas; Belin et al., 2002; Spence, 2011; Walsh, 2003), they do not produce similar patterns of results, which suggests that the underlying mechanisms that give rise to CM correspondences may not be meaningfully related to their behavior.

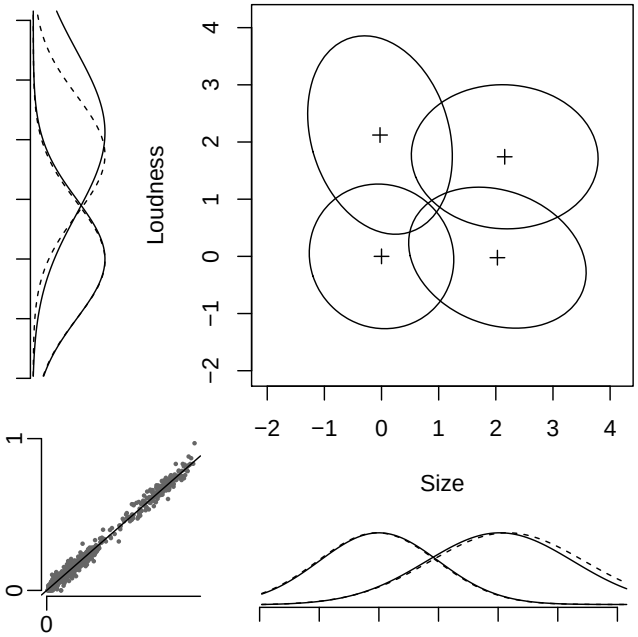

Figure 8. The results of the GRT-wIND analysis for size–loudness. In this correspondence, louder sounds are associated with larger objects and so the top right and bottom left quadrants of the graph represent the crossmodally congruent stimuli.

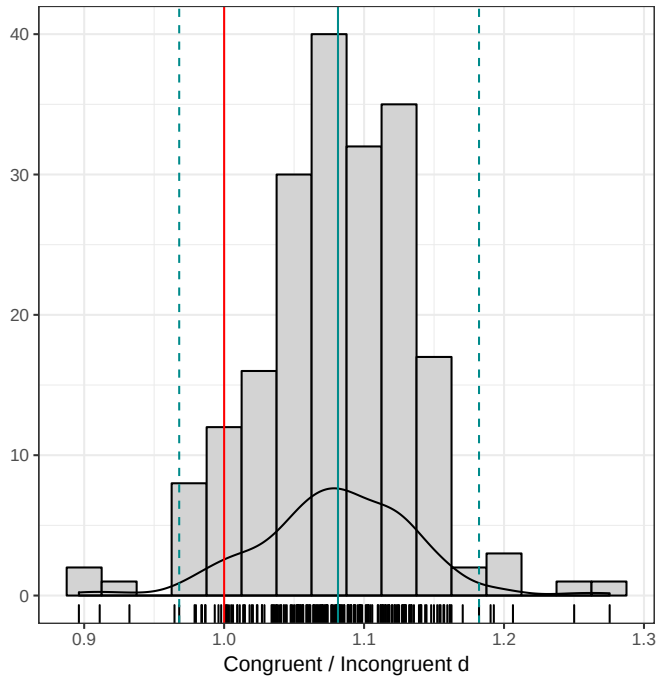

Figure 9. A diagram showing the results of the bootstrap simulation. The solid blue line shows the median congruency advantage for all simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Size–Pitch

The correspondence between visual size and auditory pitch associates larger visual objects with lower pitched tones (Evans & Treisman, 2010; Gallace & Spence, 2006). The size–pitch correspondence is likely the strongest candidate for what Spence (2011) calls a statistical correspondence, i.e., one that is born out of repeated exposure to statistical regularities in the environment. Larger physical objects tend to have lower resonance frequencies (Coward & Stevens, 2004), and so repeated exposure to this relationship is thought to have established this association in the mind (Gallace & Spence, 2006; Spence, 2011).

A similar explanation could potentially be applied to the size–loudness correspondence from the previous experiment. However, the key difference between that correspondence and size–pitch is the plausible (and empirically supported) magnitude-based explanation for the association between size and loudness (Belin et al., 2002; Spence, 2011; Walsh, 2003), which size–pitch lacks. The results of this experiment allow for a comparison to be made between these two correspondences, and highlight how environmental regularities influence crossmodal correspondences in the absence of other factors.

Methods

Participants. Data were collected from 59 participants, 11 of whom were removed for insufficient accuracy or failing to complete enough trials. After running the analysis, five individuals appeared to deviate from the rest of the sample by having decision bounds more than three MADs from the median, and were therefore excluded.

However, excluding these five participants did change the results: specifically, the likelihood ratio test indicated that the data no longer demonstrated perceptual integrality for size after removing those five participants (i.e., before removal, there was complete perceptual integrality and after removal, there is asymmetrical integrality). To keep these results consistent with those from the other experiments, this section presents results from the analysis with those five removed.

The final sample contained 43 participants: 13 men, 30 women; median age was 18.

Results and Discussion

After filtering, the median number of completed trials was 547 per participant. Median accuracy was 71.60% across all trials. Accuracy for congruent and incongruent trials were similar (69.50% for congruent trials, 73.03% for incongruent). Looking at accuracy for the two stimulus dimensions individually, they were roughly equivalent: 85.12% for size and 83.50% for pitch.

The results indicated that size–pitch had an asymmetric pattern of integrality, where size was perceptually separable and pitch was perceptually integral (the GRT model is shown in Fig. 10). Looking at the covariances of the perceptual distributions, size–pitch showed the same pattern of perceptual dependence as the other correspondences. Decisional separability failed, but only the mean decision bound in the pitch dimension had a significant slope: the mean slope was negative, indicating that participants were more biased to choose a high pitch response when the visual component was large, and more biased to make a low pitch response when the size was small.

The sensitivity analysis indicated that there was no increase in sensitivity for congruent stimuli, $d'_{ratio} = 1.06$, 95% CI: [0.94, 1.33] (results of the bootstrap simulation are shown in Fig. 11). The same was true when both dimensions were examined separately.

Size–pitch is the second stimulus pair to show an asymmetrical pattern of integrality, after brightness–loudness. These results again suggest that the mechanisms behind the correspondences have little effect on their behavior; brightness–loudness and size–loudness are both magnitude-related dimensions, but show different patterns of

perceptual integrality, as do brightness–pitch and size–pitch. Similarly, both brightness–loudness and brightness–pitch show a congruency advantage, whereas the correspondences involving size do not. Collectively, these results suggest that CM correspondences derived from statistical regularities in the environment do not show a strikingly different pattern of results than do presumably magnitude-based correspondences. The only consistent pattern in the data is in the covariances of the perceptual distributions, which again show negative correlations for incongruent compound stimuli.

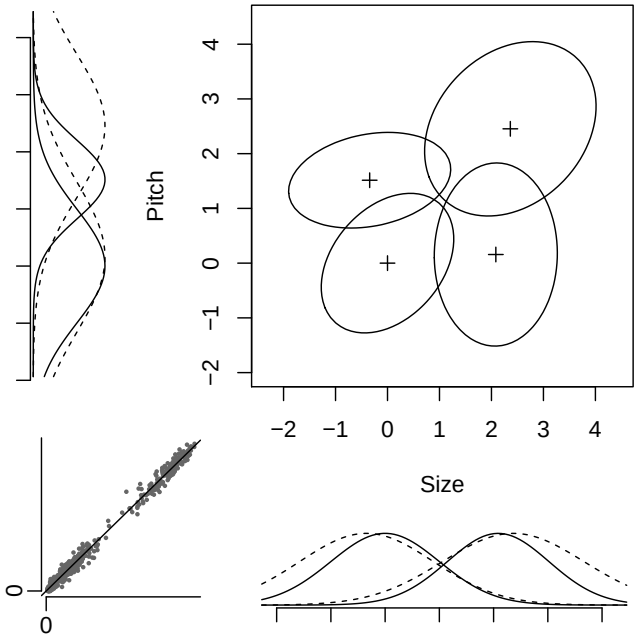

Figure 10. The results of the GRT-wIND analysis for size–pitch. Unlike the other correspondences reported on in this paper, the congruent stimuli are located in the upper left and lower right of this diagram.

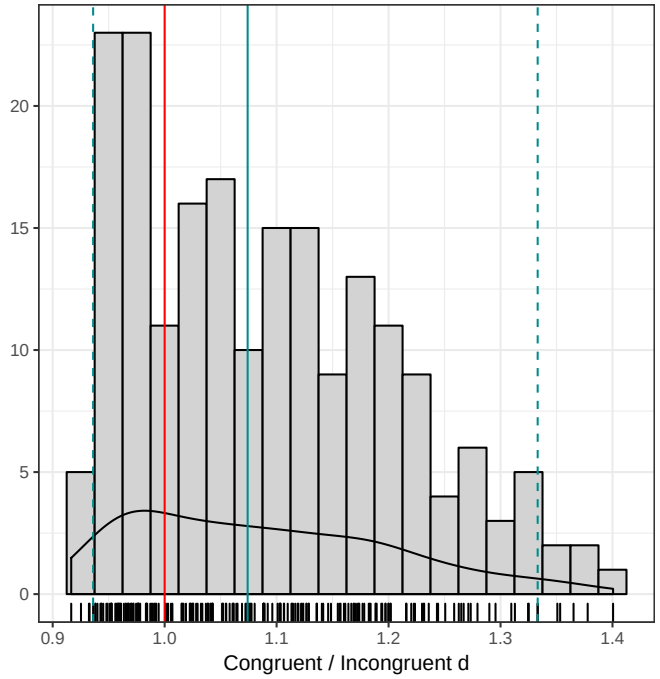

Figure 11. The distribution of the congruency advantage values from the bootstrap simulation. The solid blue line shows the median congruency advantage for all the simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Elevation–Pitch

The association between elevation (i.e., vertical position in space) and auditory pitch is one of the more well-known crossmodal correspondences in the literature, documenting a link between higher pitched tones and higher spatial locations (Bernstein & Edelstein, 1971; Evans & Treisman, 2010; Patching & Quinlan, 2002; Spence, 2011). On the surface, it would seem that this association is likely driven by the vocabulary used to describe both sets of dimensions (e.g., high places and high notes), but the association may not be purely an artifact of language. Experiments have demonstrated this association both in young children and in cultures which do not describe pitches in terms of vertical position (Parkinson, Kohler, Sievers, & Wheatley, 2012; Wagner, Winner, Cicchetti, & Gardner, 1981). While the common vocabulary may contribute to this association, it appears unlikely that this association is driven entirely by non-perceptual processes.

Methods

Participants. Data were collected from 75 individuals. Six participants were removed for not completing enough trials, as were an additional 14 for having inadequate overall accuracy. Additionally, four participants appeared to be outliers when examining the decision bounds and removed; after removing these four participants, the final sample was 51: 26 men, 25 women; median age was 19.

Results and Discussion

After filtering, the median number of completed trials was 533 per participant. For the filtered trials, median overall accuracy was 66.17%. Accuracy for congruent trials was 69.88%, and 63.30% for incongruent trials. Overall accuracy was similar for both the elevation (78.55%) and pitch (82.33%) components.

Elevation–pitch demonstrated an asymmetrical pattern of integrality, with elevation being perceptually separable and pitch being perceptually integral (the GRT model is shown in Fig. 12). The model showed perceptual dependence, again with moderate negative correlations for the incongruent stimuli. The analysis of decision bounds indicated that decisional separability did not hold for either dimension, but the mean decision bound in both dimensions was not significantly different from zero.

Analysis of perceptual sensitivity indicated that there was no advantage for congruent stimuli, $d'_{ratio} = 1.09$, 95% CI: [0.96, 1.29] (the distribution of bootstrapped congruency advantage ratios can be found in Fig. 13). An analysis the individual components showed that there was a congruency advantage for elevation, $d'_{ratio} = 1.24$, 95% CI: [1.10, 1.48], meaning that the participants had an easier time discriminating between stimuli at different elevations if they were paired with congruent tones. However, there was no significant congruency advantage for the pitch component.

Elevation–pitch demonstrates asymmetric integrality, as well as perceptual dependence, indicating that the association between elevation and pitch is not merely a semantic or otherwise non-perceptual effect. Unfortunately, the pattern of results do not offer much insight into what drives this correspondence, as the results are similar to both magnitude- and non-magnitude-based correspondences which were tested previously.

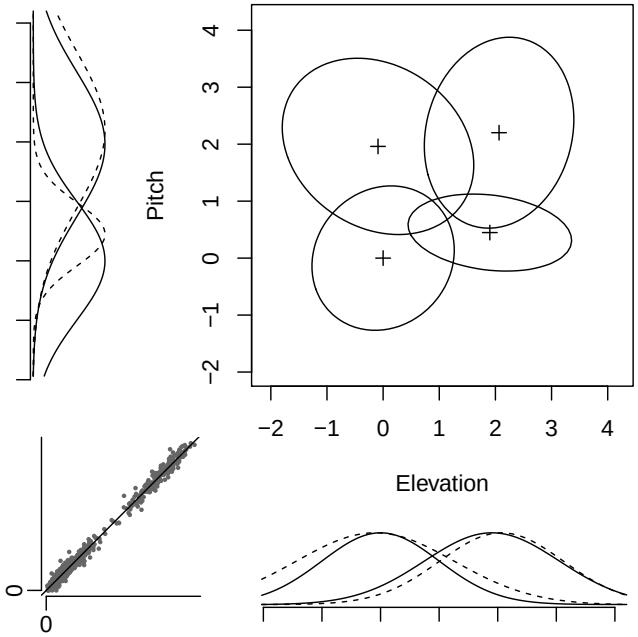

Figure 12. The results of the GRT-wIND analysis for elevation–pitch. This correspondence links high pitch with high elevation, so the crossmodally congruent compound stimuli are represented in the top right and bottom left.

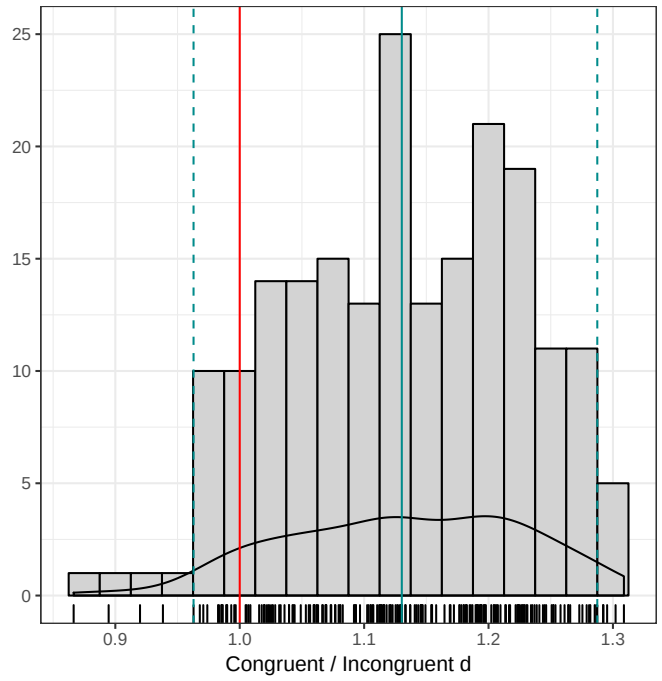

Figure 13. A diagram showing the results of the bootstrap simulation. The solid blue line shows the median congruency advantage for all simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Numerosity–Loudness

This experiment looks at the correspondence between numerosity and loudness. Although numerosity (i.e., the quantity of items in a given set; Dehaene, 1992, 2011) may seem fundamentally different from the stimulus dimensions often used in CM correspondences, it demonstrates much of the same behavior when it is part of a compound stimulus (Alards-Tomalin, Leboe-McGowan, Shaw, & Leboe-McGowan, 2014; Cohen Kadosh & Henik, 2006; Henik & Tzelgov, 1982; Oliveri et al., 2008; Szűcs & Soltész, 2008). Speeded classification tasks have found the typical congruency effect when number has been paired with both size (Henik & Tzelgov, 1982) and brightness (Cohen Kadosh & Henik, 2006); the primary difference between the numerical correspondences literature and the crossmodal literature is that the former describes the phenomenon in terms of information processing while the latter describes it in terms of perception.

This experiment tests the association between more numerous quantities and louder tones. Both dimensions can be expressed in terms of magnitude, and sharing a common format for how they are represented mentally (Walsh, 2003) may be what gives rise to this correspondence.

Methods

Participants. Data were collected from 86 participants; four participants were removed for insufficient trials, and 18 were removed for insufficient accuracy. Preliminary tests indicated that six participants had decision bound slope estimates more than three MADs from the median and were removed as outliers. The results of the analysis did not change after their removal. The final sample size was 58 subjects: 30 men, 28 women; median age was 19.

Results and Discussion

After filtering, the median number of completed trials was 527 per participant. Median overall accuracy was 61.90%. Accuracy was similar for congruent (62.15%) and incongruent trials (61.88%). Accuracy was also similar for both the number (79.38%) and loudness (76.34%) components.

The GRT model demonstrated perceptual integrality for both dimensions (the model is displayed in Fig. 14). The model also demonstrated the same pattern of perceptual dependence as was found in all previous experiments. Decisional separability did not hold for either dimension; the mean decision bound slope for numerosity was negative, indicating that there was a slight bias towards choosing a more numerous response when the tone was loud and choosing a less numerous response when the tone was soft. The mean decision bound slope for loudness was positive, which indicates that participants were slightly biased towards choosing a loud response when the number of elements in the display was low and toward choosing a soft response when the number was large.

There was a statistically significant increase in discriminability for congruent compound stimuli, $d'_{ratio} = 1.14$, 95% CI: [1.07, 1.21], as shown in Figure 15. When tested separately, neither component showed an increase in sensitivity.

This correspondence is the only one of the seven tested to show complete perceptual integrality, though it is not clear why it would be different. Both brightness–loudness and size–loudness appear to be based on magnitude processing, but show asymmetrical integrality and complete separability respectively. While both brightness–loudness and numerosity–loudness both show evidence of a congruency advantage, size–loudness does not, indicating that increases in sensitivity are not necessarily a consequence of magnitude-based correspondences. Taken together, the tests of these three correspondences strongly suggest that the mechanisms which give rise to CM correspondences do not influence their behavior.

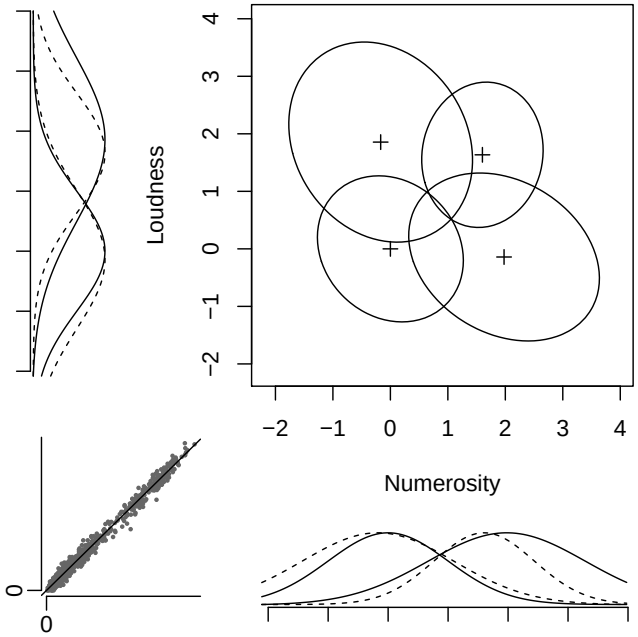

Figure 14. The results of the GRT-wIND analysis for numerosity–loudness. The congruent compound stimuli are located in the lower left and upper right.

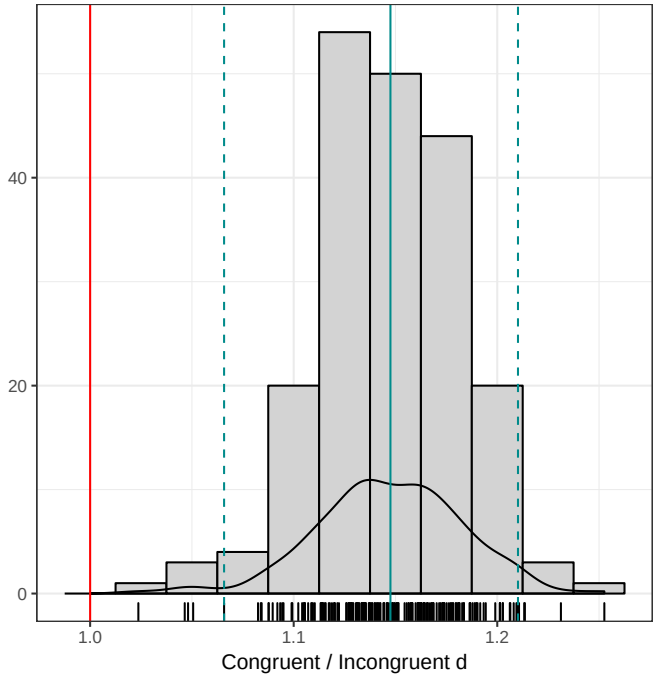

Figure 15. The median d' ratio for all bootstrap simulations. The solid blue line shows the median congruency advantage for all the simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Numerosity–Pitch

This experiment also tests an association involving numerosity, examining the association between more numerous displays and higher pitched tones (Scheid & Kubovy, in preparation). Unlike the previous experiment, there is no clear magnitude relation between the two dimensions; however, there is a linguistic connection between them, as both numbers and pitches can be described as high or low. If the difference in the mechanism that underlies this correspondence produces a difference in how these correspondences interact, it should be evident in the results of the GRT analysis.

Method

Participants. Data were collected from 64 participants. One participant was removed for not completing enough trials, 10 were removed for insufficient accuracy, and a further two were removed as outliers, resulting in a final sample of 51 participants: 20 men, 31 women; median age was 19.

Results and Discussion

After filtering, the median number of completed trials was 537 per participant. Median overall accuracy was 65.76%. Accuracy was similar for both congruent (65.26%) and incongruent trials (65.27%). Accuracy was also similar for both numerosity (79.92%) and pitch (82.77%).

The GRT model demonstrated complete perceptual separability (and is shown in Fig. 16). This model also showed perceptual dependence and the same pattern of correlations as the others. Decisional separability failed for both dimensions; the mean numerosity decision bound slope was not significantly different from zero, but the mean slope for pitch was positive, indicating a slight bias for choosing a "high pitch" response when the numerosity component was small and a bias for choosing a "low pitch" response when the numerosity component was larger.

The sensitivity analysis found no increase in sensitivity for congruent compound stimuli, $d'_{ratio} = 1.09$, 95% CI: [0.94, 1.11]; see Figure 17. Neither component of the compound stimuli demonstrated a congruency advantage.

This experiment is the second which involves numerical magnitude. Unlike numerosity–loudness, the present results do not show any evidence of perceptual integrality or an increase in sensitivity for congruent stimuli. However, it is unlikely that these differences are related to the distinction between magnitude and non-magnitude correspondences, as other non-magnitude-based correspondences have previously demonstrated some degree of perceptual integrality (e.g., brightness–pitch). Collectively, it appears that the specific stimulus dimensions involved in a correspondence have little or no relationship to what pattern of results the correspondences will show. However, despite the differing results, each experiment yielded the same pattern of covariance in the perceptual distributions; it seems likely that a negative correlation in perceptual noise for incongruent stimuli is an essential part of a crossmodal correspondence.

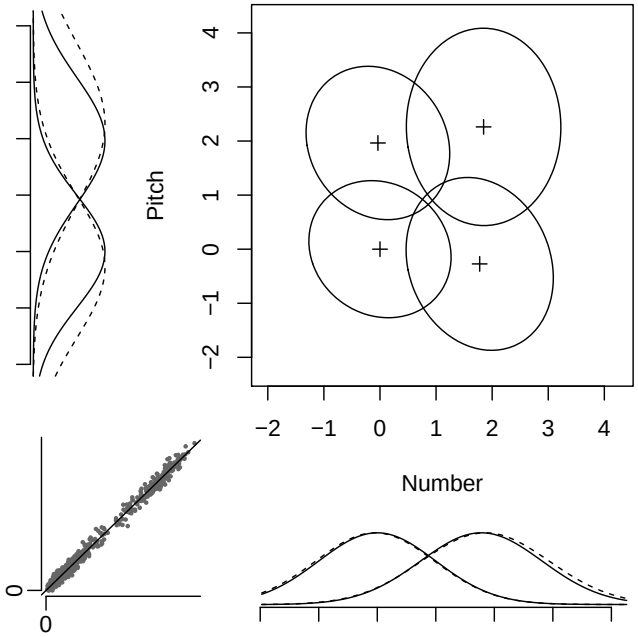

Figure 16. The results of the GRT-wIND analysis for numerosity–pitch. The congruent compound stimuli are located in the bottom left and top right.

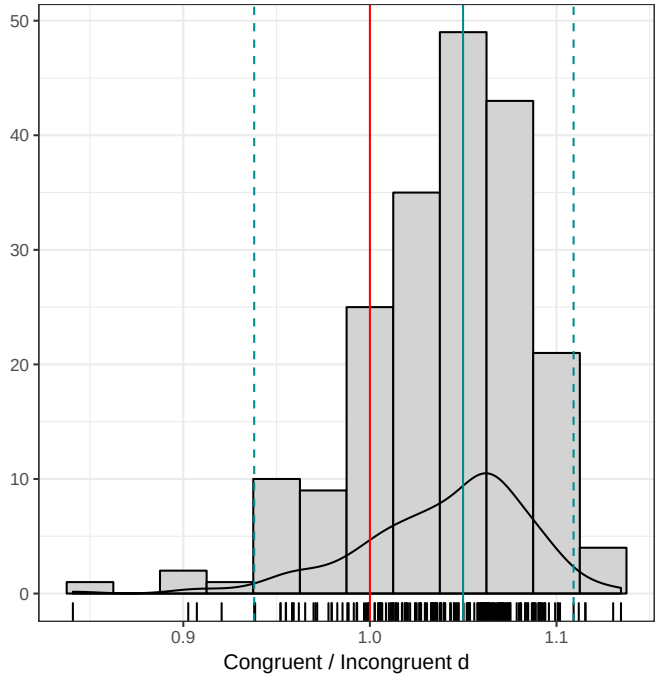

Figure 17. A diagram showing the results of the bootstrap simulation. The solid blue line shows the median congruency advantage for all simulations and can be compared to the red line, which marks the point where no congruency advantage would exist. The dashed blue lines show the 95% confidence interval around the bootstrapped median.

Stimulus Parameters

Stimulus parameters were kept consistent between experiments. The values were selected based on pilot testing to ensure that participants were able to perform well above chance while still producing a sufficient number of errors to allow their data to be analyzed (consistent with the suggestions in Soto et al., 2016).

A note on stimulus parameters: previous studies have demonstrated that CM correspondences are defined based on context (Gallace & Spence, 2006; Marks, 1987). Therefore, the absolute values of any stimulus parameter are not crucial for CM correspondences to appear; all that is required is that the two stimulus components are noticeably different on the relevant stimulus dimension. So while the particular stimulus values may or may not match what has been done in similar experiments, the findings should be compatible.

Visual Stimuli. Borrowing from Marks (1987), the brightness stimuli consisted of gray circles on a uniform black background. The brightness dimension was manipulated by changing the color value of this circle, creating the appearance of a gray circle of varying brightness suspended in a dark field. In PsychToolBox's stimulus–generating functions, color is defined by three integers which can range from 0 to 255; when all three values are the same, it produces a shade of gray with varying degrees of brightness. By changing the color values for the circle, it will appear either brighter or dimmer relative to the black background (which has the color values of 0, 0, 0). The brightness level on a given trial was chosen from one of two distributions, centered around either 45% or 70% of maximum brightness (102 and 178.5 in Matlab color values), with a standard deviation of 8.3%.

Size stimuli consisted of uniform gray circles of fixed brightness presented on a black background. The radius of the circles was manipulated to create small and large size distributions. The distributions had a mean of 40 and 60 pixels for small and large circle radii, with added jitter drawn from a Gaussian distribution with a standard deviation of 6.67 pixels. These stimuli subtended approximately 24mm (or 2.51° visual angle) and 36 mm (3.72° visual angle) across respectively, with an increase of a single standard deviation of jitter adding approximately 0.42° visual angle.

Elevation stimuli again consisted of gray circles on a black background. The circles had a fixed radius of 60 pixels; the size of the average large stimuli from the size–pitch experiment. The elevation of the stimuli was manipulated by drawing the circle some distance over or under the midline of the computer screen. The vertical position of the circle's center was drawn from two distributions that were either 20 pixels above or below the center of the screen, with a standard deviation of 13.33 pixels. Additionally, the horizontal position of the circle varied randomly between trials according to a uniform distribution so that it did not appear at the same location on each trial.

Numerosity stimuli consisted of a number of small circles placed randomly throughout the central portion of the screen (location values were limited to the middle 50% of the screen, and were selected so that no circles overlapped one another). The number of stimuli to be presented on a trial was generated in the same manner as the stimulus parameters in previous experiment, but rounded to be an integer value. The small and large number stimuli were centered around 8 and 13 with a standard deviation of 1.67. The numbers 8 and 13 were chosen so that the actual numbers used would have room to vary without being either too easy or too difficult, while falling outside the subitizing range (Dehaene, 1992, 2011; Feigenson, Dehaene, & Spelke, 2004). The circles had an average radius of 20 pixels; the size varied randomly between circles, with a standard deviation of 1 pixel. This was done so that there would not be a perfect correlation between the number of dots and the total area the dots covered on the screen.

Auditory Stimuli. For loudness stimuli, the auditory tones were 16.67 ms pure tones with a 5 ms linear on and off ramp. The pitch for each tone was fixed at 440 Hz for each trial. The two levels of volume were 64.5

and 69.4 dbSPL (equal to 40% and 70% of max volume on the computer) for soft and loud tones. The jitter was drawn from a normal distribution with a variance of 10% of max volume.[2]

For pitch stimuli, the volume was held constant, and the base frequency for low and high pitches was 330 Hz and 440 Hz. These frequencies were adjusted by a random number drawn from a normal distribution with a mean of 0 and a standard deviation of 36.67. The duration of the tones was identical to what was used for the loudness stimuli.

[2]It has been pointed out that loudness is not perceived in a linear manner, and so increasing the volume by a fixed amount will not necessarily produce the same degree of perceived change for two tones with different base levels. Therefore, if the same change in amplitude is used to create jitter on two different volume levels, it will produce unequal effects in the perceptual space. Fortunately, in a GRT analysis, this does not prevent testing for perceptual integrality/etc. — the perceptual representations for loud tones are only compared with each other, and so differences in jitter between loud and soft tones do not pose a problem.

Collected Results

The results for each of the seven experiments are presented here. Table 2 displays the results of the tests of perceptual separability/integrality, and Table 3 gives correlation coefficients for each compound stimulus from each experiment (the results of **lr_test()** indicated that perceptual independence was violated in each experiment, so only the correlation coefficients are reported). The results of the analysis of decisional separability are reported in Table 4. Table 5 has the collected results of the sensitivity analysis.

Table 2
Results of the Likelihood Ratio Test of Perceptual Separability from All Seven Experiments

Correspondence			Visual Dimension		Auditory Dimension	
Visual	Auditory	Pattern of Integrality	χ^2	p	χ^2	p
Brightness	Loudness	Asymmetrical	4.13	0.389	12.74	0.013*
Brightness	Pitch	Separable	5.26	0.261	4.09	0.394
Size	Loudness	Separable	5.33	0.255	1.18	0.881
Size	Pitch	Asymmetrical	1.48	0.831	27.60	<0.001*
Elevation	Pitch	Asymmetrical	2.18	0.702	21.02	<0.001*
Numerosity	Loudness	Integral	20.85	<0.001*	18.39	0.001*
Numerosity	Pitch	Separable	0.94	0.918	8.95	0.062

Note. A p-value smaller than 0.05 indicates a violation of perceptual separability (and is marked with an *), and identifies the stimulus dimension as being perceptually integral.

Table 3
Correlations between Stimulus Dimensions for Each Compound Stimulus from All Seven Experiments

Stimulus Dimensions		Compound Stimuli				
Visual	Auditory	A1B1	**A2B1**	**A1B2**	A2B2	mean
Brightness	Loudness	−0.01	**–0.10**	**–0.21**	0.07	−0.06
Brightness	Pitch	0.07	**–0.16**	**–0.17**	0.16	−0.03
Size	Loudness	−0.01	**–0.21**	**–0.21**	−0.02	−0.11
Size	Pitch	−0.23	**–0.26**	**–0.22**	−0.01	−0.36
Elevation	Pitch	0.04	**–0.20**	**–0.21**	0.14	−0.06
Numerosity	Loudness	−0.06	**–0.25**	**–0.18**	0.06	−0.11
Numerosity	Pitch	−0.04	**–0.16**	**–0.14**	0.00	−0.09
	mean	−0.03	**–0.19**	**–0.19**	0.06	−0.09

Note. The results of the size–pitch experiment have been inverted so that the pattern of congruent/incongruent stimuli are in line with the other experiments.
The correlation coefficients for incongruent compound stimuli are bolded.

Table 4
Results from the Likelihood Ratio Test of Decisional Separability

Correspondence		Visual			Auditory		
Visual	Auditory	χ^2	p	m	χ^2	p	m
Brightness	Loudness	159.94	< 0.001	0.04*	172.37	< 0.001	−0.09*
Brightness	Pitch	92.23	< 0.001	−0.03	67.52	0.004	−0.03
Size	Loudness	177.32	< 0.001	−0.01	106.74	< 0.001	−0.02
Size	Pitch	132.75	< 0.001	−0.01	60.49	0.040	−0.17*
Elevation	Pitch	155.37	< 0.001	0.00	103.93	< 0.001	0.02
Numerosity	Loudness	155.68	< 0.001	−0.04*	215.72	< 0.001	0.08*
Numerosity	Pitch	130.57	< 0.001	0.01	128.85	< 0.001	0.05*

Note. The m columns give the mean slope of the decision bounds. Mean slopes that are significantly different from 0 at the 0.05 level are indicated with an *.

Table 5
Sensitivity Measures from All Seven Experiments

Correspondence		d'_{ratio}		
Visual	Auditory	Overall	Visual Component	Auditory Component
Brightness	Loudness	1.23*	1.06	1.32
Brightness	Pitch	1.21*	1.19	1.22
Size	Loudness	1.08	1.06	0.91
Size	Pitch	1.06	1.07	0.67
Elevation	Pitch	1.09	1.24*	0.93
Numerosity	Loudness	1.14*	1.03	1.03
Numerosity	Pitch	1.04	0.99	0.96

Note. Values that are significant at the 0.05 level are indicated with an *.

General Discussion

CM Congruence affects Perceptual Processes

One of the goals of these experiments was to determine if crossmodal correspondences are able to influence perceptual processes, as opposed to having only non-perceptual effects, such as changing response strategies; the results indicate a relationship exists between crossmodal congruency and patterns in stimulus noise, demonstrating that crossmodal congruency does influence early perceptual processes. Each experiment showed the same trend in the covariances of the perceptual distributions: as can be seen in Table 3, while the congruent stimuli do not show any overall pattern across the seven experiments, the correlations between stimulus dimensions were uniformly negative and largely consistent in magnitude for the incongruent stimuli. This suggests that the perceptual system can recognize the lack of congruency between the stimulus components during the early stages of perception. It is unknown how or if this pattern relates to the congruency effects found in speeded classification tasks, but it is clear evidence that stimulus congruency is not solely a late-stage, decision-level process.

The results also show that perceptual integrality is not responsible for the congruency effects most associated with CM correspondences. Integrality, as defined by GRT, describes a pattern of interaction between stimulus dimensions where a perceptual representation of one stimulus dimension is influenced by the physical characteristics of both, a pattern that could plausibly produce the congruency effect found in a typical speeded classification task (Ashby & Townsend, 1986; Garner & Morton, 1969; Spence, 2011). However, the current set of results indicates that this may not be the answer. Only a few of the correspondences demonstrated some level of perceptual integrality, but all seven correspondences show a congruency effect when tested by speeded classification tasks, suggesting that another mechanism is responsible. Additionally, of the four stimulus pairs that showed perceptual integrality, three of them demonstrated asymmetrical integrality, which is at odds with results from speeded classification experiments which often show a congruency effect for both dimensions (e.g., Evans & Treisman, 2010; Marks, 1987). Collectively, these results suggest that perceptual integrality is not the driving factor behind CM correspondences.

Sensitivity Measures

The data did not consistently demonstrate improved perceptual sensitivity for congruent compound stimuli. While none of the seven stimulus pairs found a decrease in sensitivity for congruent compared to incongruent stimuli, the majority failed to demonstrate a significant perceptual advantage for congruent stimuli. Additionally, in only one case was an individual component of the stimulus pair shown to benefit from crossmodal congruency, which suggests that in complex multimodal environments where only one stimulus dimension is relevant, crossmodal congruency is unlikely to provide much benefit. This places the current results at odds with theoretical accounts that suggest CM correspondences are an adaptive mechanism (Chen & Spence, 2010; De Gelder & Bertelson, 2003; Spence, 2011); while the data indicate that it is possible for CM congruency to improve the ability to discriminate between compound stimuli, either this advantage is not universal or it is too small to be reliably detected.

The pattern of correlations between perceptual noise suggests that perceptual dependence may be the mechanism responsible for increased sensitivity for congruent stimuli. The negative covariance in the perceptual distributions for incongruent stimuli increases the degree of overlap between them, resulting in an increased likelihood of a mistaken classification. This could be responsible for the smaller d' score for incongruent stimuli even if perceptual separability holds and the amount of perceptual space between congruent and incongruent stimuli is equivalent. This would also explain why there are so few congruency advantages for the individual stimulus components, as comparing stimuli across only one dimension would make the covariances irrelevant.

Reviewing Spence's Taxonomy

Table 6
A Comparison of the Results of the Seven Experiments with the Categories of CM Correspondences Proposed by Spence (2011)

Correspondence			GRT Results		
Visual	Auditory	Category	PS	PI	Cong. Adv.
Brightness	Loudness	Struc	Asym.	No	Yes
Numerosity	Loudness	Struc	Int.	No	Yes
Numerosity	Pitch	Struc+Sem	Sep.	No	No
Brightness	Pitch	Struc+Sem (?)	Sep.	No	Yes
Size	Loudness	Struc+Stat	Sep.	No	No
Size	Pitch	Stat	Asym.	No	No
Elevation	Pitch	Stat+Sem	Asym.	No	Elevation only

Note. Spence's three categories are abbreviated as **Struc**tural, **Stat**istical, and **Sem**antic.
Brightness–pitch is labelled with a (?) because Spence (2011) was uncertain as to how to classify it. The classifications for the two numerosity experiments are extrapolated based on the description of the relevant categories.
The Cong. Adv. column indicates if a significant congruency advantage was found for the compound stimuli.
The elevation–pitch correspondence found a significant congruency advantage only for the elevation component.

In Table 6, the results of the seven GRT experiments are displayed alongside the categories of crossmodal correspondences proposed by Spence (2011). Spence's categories do not appear to match patterns in the empirical results; neither the pattern of perceptual integrality nor the presence of a congruency advantage follow the distinctions that Spence draws between categories of correspondences.

Of Spence's three categories, structural correspondences have the strongest support in the literature, as this category encompasses the magnitude-based associations described by Walsh's "A Theory of Magnitude" (Walsh, 2003). However, the present data cast doubt that Spence's categories can make predictions about how correspondences behave. The three best examples of magnitude-based correspondences that would fall under the structural category (size–loudness, brightness–loudness, and numerosity–loudness) all show different patterns of perceptual integrality, and they do not consistently or exclusively show a congruency advantage. And so, while Spence's taxonomy is good at highlighting similarities between different crossmodal correspondences, it appears unable to make accurate predictions about the characteristics of these correspondences based on these similarities.

Limitations

One limitation of these experiments is that the analyses cannot separate the variance attributable to the stimulus jitter from the variance related to perceptual noise. While the amount of physical jitter is recorded when it is added to the stimulus, there is no method to translate the different physical stimulus values into values that are scaled for perceptual space. However, the way in which the stimulus jitter affected the results can be anticipated;

because the degree of jitter was identical for stimulus components of the same value (e.g., in the brightness–loudness experiment, both bright/soft and bright/loud stimuli had the same degree of jitter applied to the brightness component), the variance both perceptual distributions would share an identical component, which would have the effect of making their overall variances more similar. In terms of the results of the GRT analysis, if the stimulus pair had variance-based perceptual integrality, it would be less evident due to the inclusion of the stimulus jitter.

It is unlikely that this effect influenced the overall results from the experiments. The jitter for each set of stimulus values was generated in an identical manner: all jitter values were drawn from normal Gaussian distributions with standard deviations equivalent to one-third the difference between the two mean stimulus values, meaning that the effect of jitter would be roughly similar for all stimulus dimensions in terms of added variance. Yet, multiple experiments detected perceptual integrality, including cases where it appeared that the effect was largely driven by differences in perceptual variances, such as the brightness–loudness experiment. This suggests that while the presence of stimulus jitter undoubtedly influenced the estimated GRT model, it was not strong enough to obscure the true patterns of integrality.

The inclusion of stimulus jitter affected the results in a second way. Because the jitter was randomly generated on each trial, this meant that there was a source of variance in each dimension that was uncorrelated with one another. This uncorrelated variance has the effect of decreasing the overall covariances for each of the perceptual distributions, which in turn makes the GRT analysis more likely to indicate that perceptual independence holds. However, given that the results indicated a violation of perceptual independence in every experiment, it is likely that the jitter's influence on the perceptual dimensions' covariances was minimal.

Open Questions

Some open questions remain. First, there is the question of how the results of the GRT analyses connect to the congruency effects found in speeded classification tasks. Congruency effects are consistently found in these tasks, but the only consistent finding in the GRT results was the pattern of covariances found in each experiment, suggesting that it is an essential feature of the correspondence. However, it is unclear as to how this pattern could produce the congruency effects found in speeded classification tasks, or whether this mechanism is responsible for the congruency effect at all. A non-zero covariance in a perceptual distribution would influence the type and frequency of mistakes made in a GRT task where participants are asked to make a discrimination on two different stimulus dimensions simultaneously (i.e., if the perceptual distribution in the upper left corner had a strongly negative covariance, the distribution would stretch diagonally towards the point where the decision bounds intersect, making trials where both stimulus components were misclassified more common than if the covariance was zero), but would not influence behavior when only a single discrimination is made, provided the participant uses an unbiased decision rule. A non-zero covariance could affect behavior in a single discrimination task (like the speeded classification task) if a biased decision rule was followed, but unfortunately this hypothesis cannot be tested with the data at hand.

Secondly, it is not clear what factors are responsible for the differing patterns of perceptual integrality and congruency advantages between the different correspondences. Generally, all CM correspondences appear largely the same in speeded classification tasks, but this may be attributable to the narrow scope of data they provide. Congruency effects measured in RT are troublesome as they cannot be easily compared across different stimulus pairs as not all stimulus dimensions are necessarily perceived at the same speed. Additionally, these tasks often cannot state whether a congruency effect represents a facilitation effect for congruent stimuli or an interference effect for incongruent stimuli, as baseline trials are often either faster or slower than either (Evans & Treisman, 2010; Gallace & Spence, 2006; Sanabria et al., 2007, etc.). The limitations of the data provided by speeded classification tasks was a part of the impetus for moving to GRT, so inconsistencies between previous literature

and the current set of results were expected. However, while the richer data set from current set of experiments has been able to uncover new divisions between different CM correspondences, it cannot yet explain why these differences exist.

Conclusion

By adopting the GRT paradigm, the current experiments have been able to shed light on several questions about crossmodal correspondences. The results of the seven experiments show that crossmodal correspondences do exert an effect at the perceptual level, though decision-level processes are likely affected as well. The only finding common to all the CM correspondences is a pattern of correlation between stimulus noise for the incongruent stimuli, which demonstrates a perception-level effect of stimulus congruency, although it is unclear how or if this effect is related to the congruency effects from speeded classification tasks. Additionally, the data suggest that crossmodal correspondences do not consistently produce an increase in perceptual sensitivity for congruent stimuli. While some correspondences did show a significant increase in discriminability for congruent stimuli, which shows partial support for theories that suggest CM correspondences have ecological value (Chen & Spence, 2010; De Gelder & Bertelson, 2003), this does not appear to a universal characteristic of crossmodal associations.

Additionally, the results of these experiments allow for an evaluation of the classification scheme for CM correspondences devised by Spence (2011). His taxonomy is the most comprehensive attempt to date to organize the wide array of different crossmodal correspondences into a theoretical structure. However, the divisions he specified do not match patterns in the experimental results, which suggests that the taxonomy's categories may not capture meaningful differences in how CM correspondences behave. A large-scale revision of the taxonomy may be required to incorporate the results of the current set of experiments with earlier research.

Appendix: Introduction to the Theory and Methods of GRT

This section is designed to give a general overview of the experimental design and analysis of GRT experiments and provide sufficient information for a reader to understand and evaluate the results.

Before describing the mechanics of GRT, it would be useful to explicitly define some of the terminology used to describe stimuli and their perceptual representations. The terms "physical stimulus" and "physical stimulus level" or "physical stimulus value" refer to the stimulus itself (i.e., the image or tone displayed to the participant) and the particular parameters used to create that stimulus (such as the particular size in pixels of the visual component or the frequency in Hz of the auditory component). Conversely, the term "perceptual representation" refers to the internal representation of the compound stimulus. GRT models the internal representation of a compound stimulus as a point on a two-dimensional plane: the two axes of the plane represent the two stimulus dimensions and the "perceptual level" or "perceptual value" of the stimulus on each dimension determines the point on that plane where the perceptual representation falls.

This explanation will focus on a hypothetical 2×2 study in which the stimulus dimensions being tested are the color (either light or dark gray) and the width of a box (such that it will be either a square or a rectangle). A representation of GRT space for this example is depicted in Figure 18.

The purpose of GRT is to model the internal processes and variables that take place between stimulus and response when multiple stimulus dimensions are presented simultaneously (Ashby & Townsend, 1986). The GRT model describes this process in two stages: the first describes how physical levels of stimuli are translated into perceptual values within the perceptual system, and the second describes the internal rules that determine which perceptual values are mapped to which response options. Because GRT is able to model both processes, it allows for researchers to produce separate estimates of how the two stimulus dimensions interact at both the perceptual

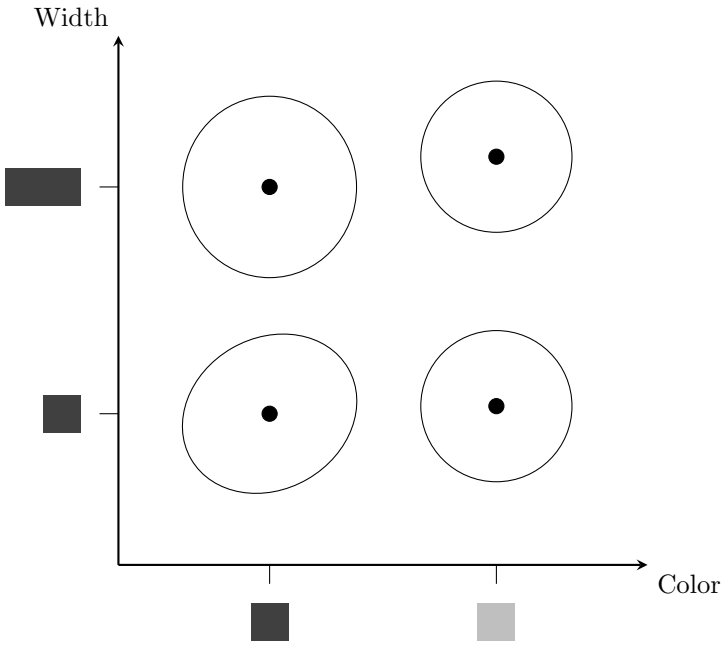

Figure 18. A hypothetical perceptual space from an experiment looking at the perception of color (in gray-scale) and box width. Each compound stimulus is represented as a bivariate normal distribution.

and decision level. As a result, GRT experiments are useful for explicitly separating perceptual results from potentially less interesting decision-level effects.

Because GRT is built upon Signal Detection Theory (SDT), it retains the same fundamental assumptions about perception. The first of these is the assertion that all perception is noisy (Ashby & Soto, 2014); the internal perceptual representation of a stimulus is not determined solely by its physical value, but is also influenced by random noise (both external physical noise and internal neurological noise). Because of this, the process that translates the parameters of the physical stimulus to a perceptual representation will produce a range of different results after repeated presentations of the same stimulus. The noise is typically assumed to be random and normally distributed, which means that in our example, a particular physical value for the box width component will produce a normally distributed distribution of possible perceptual width values. When the amount of noise becomes larger, the variance of the distribution of perceptual values increases, which has the effect of increasing the degree of overlap between adjacent distributions along the width dimension. This increase in overlap leads to an increase in the number of errors when participants are tasked with discriminating between the two levels of that dimension.

The second borrowed assumption is the idea that responses are selected by a separate decision process that divides the perceptual space into distinct regions, each of which is associated with a particular response. In a 2×2 GRT analysis, perceptual space is modeled as a two dimensional plane with two border lines (often called decision bounds) which divide that space into four discrete regions, each of which is associated with one of the four possible responses. The perceptual process translates the external physical stimulus into a internal representation consisting of a point on this 2D plane, and where this point falls relative to the decision bounds determines which response the participant makes. Because the perceptual process includes random noise, some stimuli will be misclassified as their perceptual representations fall on the wrong side of a decision bound; the amount of variance associated with a stimulus, as well as the distance between the center of the stimulus's distribution and the decision bound, determine the likelihood of this occurring on any given trial.

Questions Which GRT Can Address

The model generated by a GRT experiment can be used to answer three main questions about how the stimulus dimensions interact (in practice, not all of the questions are necessarily answerable with the available data; this topic will be covered in greater detail in a later section). The first two questions, perceptual separability-vs-integrality and perceptual independence-vs-dependence, describe how physical stimuli are transformed into internal perceptual representations. The third is related to the decision-level process that maps these internal representations to response categories.

The distinction between perceptual separability and perceptual integrality refers to whether the perception of one stimulus dimension is influenced by the other, in a way that changes either the mean perceptual value or the variability of those perceptual values over repeated presentations. Returning to the color–width example, if perceptual separability holds, an observer's perceptual representation of a given stimulus dimension, e.g., how "dark" the shape was perceived to be, would be influenced solely by the physical level of the color (and the random perceptual noise on that particular trial). Figure 19 illustrates how the color dimension in the example experiment demonstrates perceptual separability: the perceptual distributions, both in mean and variance, do not change as a function of box width (if you compare any of the four perceptual distributions with the one either above or below it, both will have their mean located at the same point on the x-axis, as well as equal degrees of variance).

In contrast, perceptual integrality is when the two stimulus dimensions are not processed independently and the perception of one dimension is influenced by the physical level of both stimulus components. In the color–width example, the width dimension shows perceptual integrality (illustrated in Figure 20); the perceived level

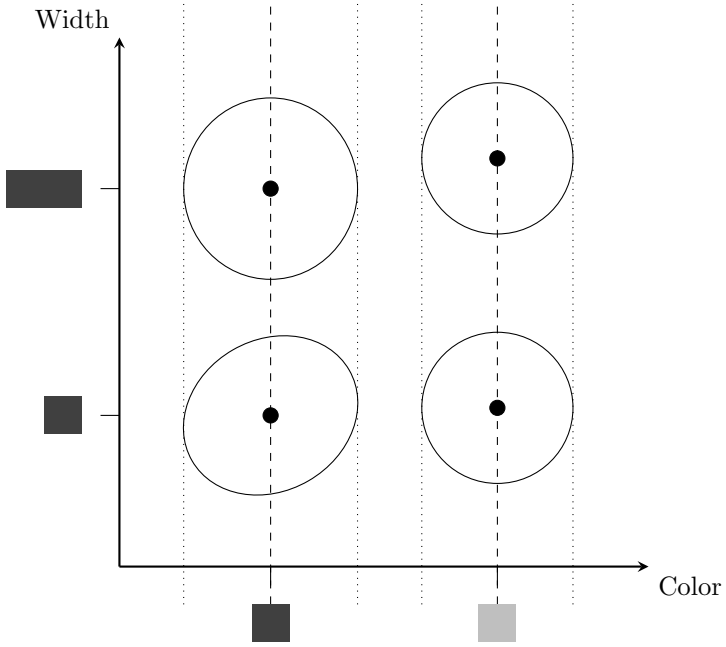

Figure 19. An illustration of how perceptual separability appears in a GRT plot. Both the means and variances in the color dimension remain the same despite changes in the other stimulus dimension.

of width changes depending on the level of color. Comparing the dark/wide and light/wide compound stimuli, the perceived width of the stimulus is greater (i.e., higher on the y-axis) when an identical rectangle is drawn in light colors as opposed to dark. Perceptual separability does not have to be symmetrical; as in this example, one dimension can be separable from the second, while the second is integral with the first.

The second question that GRT can answer is perceptual independence-vs-dependence, which refers to the statistical independence (or lack thereof) between the perceptual noise on each stimulus dimension. As mentioned previously, GRT assumes that there is some perceptual noise on each trial, meaning that subsequent presentations of identical physical stimuli will produce a distribution of perceptual values; when the noise values on both stimulus dimensions are statistically independent, then the stimulus dimensions show perceptual independence. A significant correlation between the dimensions is indicative of a perceptual-level interaction between the stimulus dimensions, suggesting that there is a source of noise that is common to both dimensions. Unlike perceptual separability, independence is a characteristic of a particular compound stimulus, and not something that describes the relationship between stimuli, and as a result, may hold for some stimuli and not others (for example, in Fig. 18, there is evidence of perceptual dependence for the stimulus distribution in the bottom left, but not for the other stimuli).

The third question relates to decisional separability-vs-integrality, which describes the relationship between the decision bounds and the stimulus dimensions with which they are associated. In a 2×2 GRT experiment, each stimulus dimension has a decision bound, which represents the boundary line between one response option and the other for that dimension, e.g., separating a "wide" width response from a "narrow" response. If decisional separability holds, as it does for width in Figure 21, it means that only the perceived value of the relevant dimension influences how a stimulus will be categorized, e.g., only the perception of the stimulus's width is used to

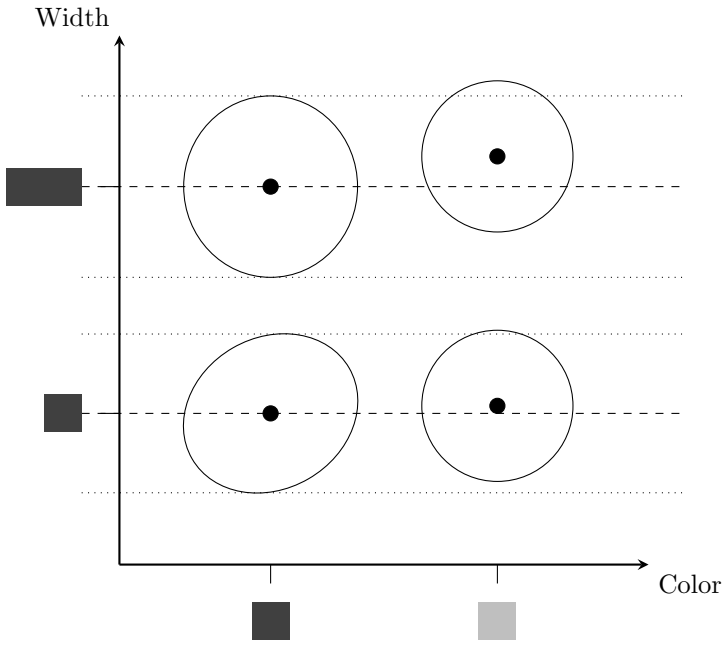

Figure 20. An illustration of how perceptual integrality appears in a GRT plot. Width demonstrates perceptual integrality because the perceptual representations of width are affected by the physical levels of both stimulus dimensions.

determine whether a "wide" or "narrow" response is made. Conversely, if perceptual integrality is demonstrated, as is the case for color, then the criteria for separating one response from another varies as a function of the other dimension, e.g., the amount of perceived lightness that is needed to make a "light" response over "dark" changes depending on the perceived level of width.

Whether an individual displays decisional separability or integrality is largely dependent on the nature of the experimental task (Ashby & Townsend, 1986; Soto et al., 2015). Decisional separability is often expected in a GRT experiment, because the stimulus values on each trial are chosen randomly and known to be uncorrelated with one another. However, if stimulus values are correlated, a decisionally integral response strategy can become optimal — if the participant knew that squares were almost always shown with dark colors, he or she may be more likely to classify a shape as a square if it was displayed in the dark color.

All three of these characteristics (Perceptual Separability, Perceptual Independence, and Decisional Separability) are all separate elements of a GRT model. A particular pair of dimensions may show any combination of these three characteristics.

Experimental Design

The prototypical GRT experiment involves participants classifying four different compound stimuli, which were created by combining two stimulus dimensions with two levels each. In the color–width example, the four compound stimuli would be a dark square, a light square, a dark rectangle and a light rectangle. On each trial, the participant is presented with a randomly chosen compound stimuli and asked to categorize it as one of the four types. The data collected in the experiment are a count (or proportion) of how many times each of the four

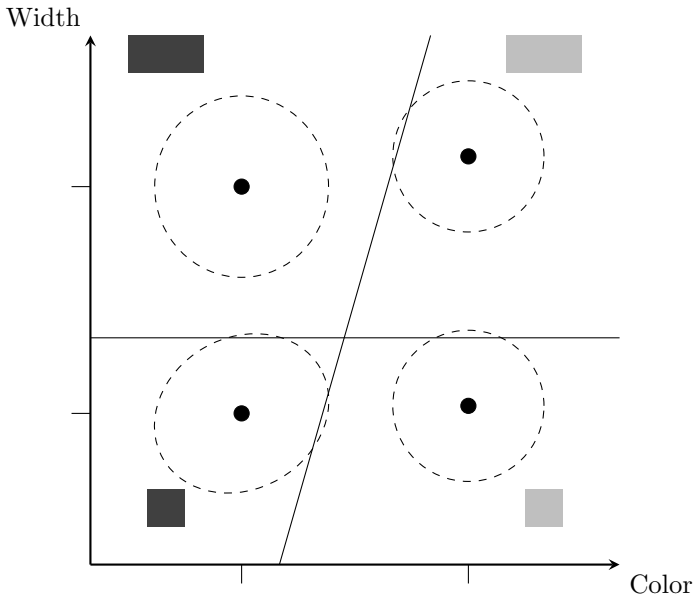

Figure 21. An illustration depicting a pair of hypothetical decision bounds along with the four responses associated with each region. The width decision bound displays decisional separability (as it is perpendicular to its associated axis) while the color decision bound shows decisional integrality.

compound stimuli garnered each response. The data are typically assembled into a confusion matrix, to allow for an easy comparison of how many incorrect responses a participant made to a particular stimulus and which type of error it was.

Once the data have been collected, they may be analyzed separately to produce a separate model for each individual participant or multiple participants can be analyzed together; separate R packages are available for either type of analysis, e.g., the **mdsdt** package, (Hawkins, Houpt, Silbert, Blaha, & Wickens, 2016) and the **grtools** package, (Soto et al., 2016) respectively. In either case, the data are used to estimate various parameters (detailed below) and describe perceptual and decisional characteristics of the interaction between the two dimensions.

Parameters of a GRT Model

A hypothetical GRT model for a 2×2 experiment is defined by 16 parameters related to perceptual separability/integrality, four related to perceptual independence, and four related to decisional separability, though not all of them are necessarily able to be estimated in any given analysis. These parameters describe both the characteristics of the perceptual distributions in perceptual space (the first 20) as well as the decision rules used to associate a set of perceptual values with a particular response (the last four). The representation of perceptual space takes the form of a two-dimensional plane, with each axis representing one of the two dimensions. The particular numerical values given to this space are arbitrary; typically, the centers of one or more perceptual distributions will be fixed in order to be used as a reference point for other parameters, and so the particular values of these parameters are only meaningful in relation to others on the same dimension. Additionally, there is often no theoretical guidance for scaling distance on the two axes, so typically no attempt is made to do so (as determining what change in color

is equivalent to a one centimeter increase in box width is not the goal of a GRT experiment). Fortunately, none of the perceptual or decisional characteristics require a comparison between stimulus dimensions in order to be measured, so equating the dimensions is not necessary.

Parameters related to Perceptual Separability/Integrality. As mentioned previously, the process of perceiving a stimulus and selecting a response can be divided into two stages: perceptual and response. At the perceptual level, each compound stimulus is represented by a (typically assumed to be normal) bivariate distribution defined by two parameters for the distribution's center along each axis and two parameters for the variance in each axis (there is also a fifth parameter describes the covariance of the distribution; this will be covered in the section on perceptual independence).

The parameters related to the center of the distribution describe the mean perceived value on each dimension for the corresponding compound stimulus. As one would expect, two stimuli with different physical color values would be located in different positions with regard to the x-axis. The important comparison is between two stimuli with identical color values, but different width values; if the mean x value is the same for both levels of width, then that is consistent with perceptual separability. If the two x values differ from one another despite the physical level of color being identical, it indicates that the perception of color is influenced by both the stimulus's color and width, and demonstrates perceptual integrality.

The variance parameters describe the degree of variability in perceptual representations that is due to perceptual noise for a particular stimulus. The variance in the x dimension describes the amount of variability that exists in the perception of the color component. When comparing two different compound stimuli that differ only in the color component, there is no expectation that both levels of color would have the same amount of x dimension variance, and none of the standard GRT tests involve a direct comparison of that type. As with the means, a test for perceptual separability would involve comparing the amount of perceptual variance along the x-axis for two stimuli with the same level of color but with different width levels (see Fig. 19). Perceptual separability is indicated by both the means and the variances remaining stable across different levels of the other dimension, meaning that the perception of color is completely unaffected by the level of width; if one or both of those criteria are violated (as they are in Fig. 20), then the stimulus dimensions demonstrates perceptual integrality.

Perceptual Independence/Dependence. Testing for perceptual independence involves estimating the covariances for each perceptual distribution. If perceptual independence holds, then the levels of perceptual noise present in each stimulus dimension will be uncorrelated and the resulting stimulus distribution will have a covariance of zero. In Figure 18, three of the four stimuli display perceptual independence (all but the dark/narrow stimulus in the lower left). Conversely, perceptual independence does not hold for the dark/narrow stimulus; the positive covariance indicates that when the width component of one of these stimuli is perceived as being relatively more wide, the color is perceived to be relatively brighter.

Decisional Separability/Integrality. A typical 2×2 GRT experiment estimates four parameters related to decisional separability: a slope and intercept for both decision bounds. As a 2×2 GRT experiment only has 4 total stimuli, the decision bounds are typically assumed to be straight lines. There is no theoretical reason that the decision bounds must be straight (see Ashby & Maddox, 1990); however, estimating the decision bounds in more detail would require a greater number of stimuli in the experiment, which is often impractical.

Of the four decision bound parameters that are estimated, typically the two slope parameters are the ones of interest. The intercept parameters indicate overall biases (i.e., whether an individual is more or less likely to give a "wide" response compared to a "narrow" response in general), but the slope parameters determine whether the individual displays decisional separability or integrality (however, decision bounds cannot always be estimated in a GRT analysis; this issue is covered in a later section).

When decisional separability holds, it describes a situation in which the individual's internal decision rule that marks the boundary between one response and another is unaffected by the level of the other stimulus dimension;

this is illustrated by the decision bound for width in Figure 21. Decisional separability is marked by a slope parameter equal to zero: the decision bound is perpendicular to the main axis, and indicates that a given perceived value of width will always produce the same size response; no matter how light or dark the box is perceived to be, the response for the width component would not change.

Conversely, the decision bound for the color dimension has a non-zero slope, which indicates decisional integrality. This describes a situation where the perceptual values of both stimulus dimensions determine how the participant responds to color. In this example, the participant would be relatively more likely to respond 'light" overall when the stimulus is narrow, and more likely to respond "dark" when the stimulus is wide.

Both decisional separability and integrality can represent optimal decision strategies under different conditions. When the stimuli are chosen randomly, and there is equal probability of any combination of stimuli appearing on a trial, decisional separability is ideal. However, if the stimulus values were chosen so that stimuli were more likely to be wide and dark or thin and light, and less likely to be wide and light or thin and dark, then non-perpendicular decision bounds would be better.

Types of GRT Analyses

A typical GRT analysis is carried out on a single participant's data. However, it is not possible to estimate all the parameters involved in a 2×2 GRT model with only one confusion matrix — the confusion matrix would have 16 cells, which give 12 degrees of freedom, and a full GRT model has 20 free parameters: two means and two variances for four distributions, four covariances, and four values related to decision bounds. There are two main ways to get around this limitation. The first way is most common, and involves fixing enough parameters so that the data are sufficient to get at the information of interest. The second combines data from multiple participants to have sufficient degrees of freedom to estimate all the parameters of interest in a single common model (GRT-wIND, Soto et al., 2015).

Standard GRT Analysis. The traditional analysis is what has been used to investigate perceptual separability for the majority of GRT experiments. To get around the limitations of the data, a few of the different types of parameters are fixed so that the parameters of greater theoretical interest can be estimated. In a typical experiment, the parameters related to decision bounds are fixed, constraining the model so that decisional separability must hold. Assessing decisional separability is often of less importance than the perceptual parameters, and given that most experiments use stimulus values that are uncorrelated with one another, it is a plausible assumption. Additionally, either the means or the variances of the distributions are free while the other is fixed (with a single pair of decision bounds, changing the center of the distribution is equivalent to changing the variance). With these parameters fixed, the analysis cannot assess decisional separability, but is capable of assessing perceptual separability; if perceptual integrality exists, regardless of whether the true effect primarily occurs in a difference in mean values or a difference in variance, it will be detectable as the effects will appear in whichever of the two sets of parameters remain free in the model. The standard GRT analysis also allows for estimating the covariances of the distribution, so perceptual independence can be examined as well.

As mentioned above, the major limitation of the standard GRT analysis is its inability to estimate decision bounds. Because the decision bounds are fixed to be perpendicular, it imposes a restriction on how the model is estimated. In a completely free model, all of the parameters can be rotated together without producing any drop in model fit. In a hypothetical model, the center of the first distribution is typically fixed at (0, 0) or (1, 1), and rotated so that the pattern of the distributions makes theoretical sense (i.e., two stimuli that only differ on one stimulus dimension will be generally aligned along that dimension), but the model could be rotated arbitrarily around the center of the first distribution and still fit the data equally well. This means that if decisional separability does not truly hold, then the true GRT model will be distorted when transformed to meet the assumption of decisional

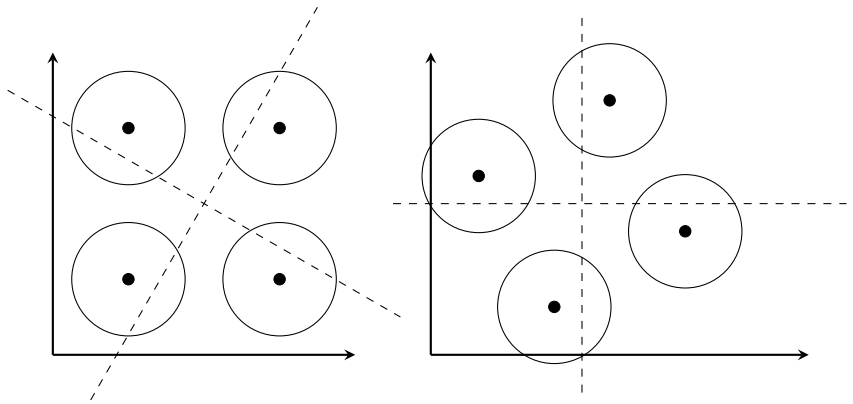

Figure 22. An illustration depicting the potential for distortion when rotating a hypothetical GRT model to ensure decisional separability. The left shows the true GRT model, with perceptual separability and decisional integrality, and the right shows the model after being rotated to show decisional separability (and which now shows perceptual integrality). Both models fit the data equally well, but have different theoretical interpretations.

separability imposed by the standard GRT analysis. A simplified example is illustrated in Figure 22; the true GRT model on the left shows perceptual separability and decisional integrality. However, when rotated in order to establish decisional separability, the locations of the perceptual distributions are altered and the model no longer demonstrates perceptual separability. This is a fairly benign example, as the decision bounds were perpendicular to each other before and after the rotation; however, if the angle between the true decision bounds was not equal to 90°, then the amount of distortion in the rotated model would be even greater.

GRT with Individual Differences. An alternative to the standard analysis is GRT with Individual Differences (GRT-wIND; Soto et al., 2015). GRT-wIND was devised to get around the problem of wanting to estimate more parameters than there are degrees of freedom in a single confusion matrix by analyzing multiple individuals together. The additional data allows for the estimation of both parameters related to the perceptual distributions and those related to decision bounds. The analysis is based on the assumption that perceptual processes are the same between individuals, and thus perceptual separability and independence are the same across participants, while decision bounds can be different for each person. This means the analysis can estimate a common set of both the means and variance parameters for the distributions (with the exception of the first distribution, which is fixed to be centered at (0, 0) with a variance of 1 for both dimensions, for the other distributions to use as a reference), as well as their covariance, for the sample as a whole, in addition to estimating a set of decision bounds for each individual participant.

GRT-wIND gets its name from the two sets of individual difference parameters that it estimates. The first set is a pair of decision bounds estimated for each participant. Giving each participant their own decision bounds not only allows the model the better fit each participant's data, it also allows for decisional separability to be evaluated. When multiple decision bounds are present in the model, it prevents any single rotation of the model from creating the conditions for decisional separability for all participants (with the exception of when all sets of decision bounds are parallel to one another). Because of this, decisional separability can be evaluated by comparing the unrestricted model to one where all decision bounds are fixed to have a slope of 0. If the constrained model does not have

worse fit, it indicates that all the decision bounds can be made perpendicular to the relevant axis while fitting the data just as well.[3]

The other set of parameters which GRT-wIND estimates for each participant are κ and λ. These parameters allow the model to better account for differences in performance between participants by rescaling the perceptual distributions relative to each other in a way that does not alter their overall structure in perceptual space. κ scales the perceptual space by shrinking the variances for the distributions in both dimensions equally, so as to account for differences in individuals' overall performance, while λ selectively shrinks the variances of one dimension relative to the other, to account for individuals who do not perform equally well on both dimensions. These parameters are described in terms of attention, though this term was chosen more for convenience than any theoretical reason; their main function is to adjust the variances of the distributions to better account for an individual's performance. κ values are described as measuring global attention, as larger κ values produce smaller variances, and therefore less overlap between adjacent distributions, and is used to calibrate for the participant's overall level of performance. λ is described as selective attention; it has a value between zero and one and describes the participant performing better on one dimension over the other. Values greater than 0.5 shrink the variances on the x dimension while increasing those on the y dimension (indicating better performance on the x dimension compared to the y).

The **grtools** package carries out five different tests when evaluating a GRT model: testing perceptual separability and decisional separability for both dimensions individually, and perceptual independence for all four compound stimuli collectively. This presents something of a problem, as illustrated by the hypothetical GRT model shown in Figure 23, which displays three GRT models that would each fit a set of data equally well. The only constraints to the model that the GRT-wIND analysis imposes are that the first distribution has a center at a specified point with a variance of one in both dimensions, which allows for the entire model to be rotated around that center point without any loss of model fit. This allows different rotations of the same model to demonstrate perceptual separability for the Y dimension, perceptual separability for the X dimension, or neither, without any loss in model fit (as is shown in Fig 23). The problem is that when our hypothetical data are tested for perceptual separability, only one constraint is fit at a time, so the unconstrained model would be compared to the restricted model on the left when testing for perceptual separability for the Y dimension (and find no loss of fit) and to the center model when testing for perceptual separability for the X dimension (again finding no loss of fit). Because

[3]However, because the whole GRT model can be rotated freely without a loss of model fit, the results of this test are best interpreted as indicating whether all the decision bounds from each dimension are essentially parallel to each other and the actual slope of the lines needs to be assessed separately.

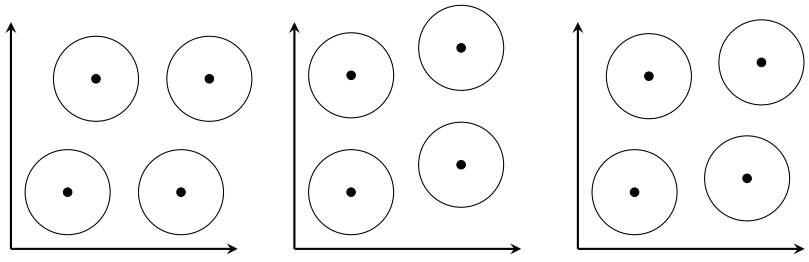

Figure 23. Three possible GRT representations that each fit a hypothetical set of data equally well. Depending on which rotation of the model is chosen, perceptual separability will hold for either the stimulus dimension on the y-axis, the x-axis, or neither.

all of these models are simple rotations of one another, each will fit equally well, and perceptual separability would be indicated for both dimensions when tested individually. However, if this model were tested for perceptual separability in both dimensions simultaneously, the results would indicate that the data does not support this constraint. The only conclusion that could be drawn with these results is that both dimensions are not separable, but that either of the two dimensions could show perceptual separability, though we cannot know which one, if either.

A similar problem exists for testing perceptual independence. If any distribution has a non-zero covariance, it is possible to rotate the entire model so that particular distribution is made to have zero covariance with no reduction in model fit. However, if the perceptual distributions do not all have the same covariance, there is no rotation that would allow all distributions to show perceptual independence, and for this reason, the test of perceptual independence is best understood as a test for whether or not experiment-wide perceptual independence could exist.

A traditional GRT analysis makes further comparisons, comparing the unrestricted model to one where perceptual separability was fixed in both dimensions simultaneously, among others. However, this would not solve the problem of rotation mentioned above, as the results could indicate either that perceptual separability holds for both dimensions, or that it does not, and in a case like the one in Figure 23, it would still not allow for any of the three possibilities to be ruled out. The fact that additional tests do not provide much added utility and would be computationally expensive (a traditional GRT analysis runs 11 different models, testing different combinations of perceptual separability and perceptual independence, but including decisional separability or testing any of the individual covariances for perceptual independence). It is likely for this reason that the **grtools** package only runs the five tests mentioned above.

Sensitivity Measures

As mentioned earlier, participants' sensitivity to differences in compound stimuli can be measured through a multivariate adaptation of SDT's d' statistic. d' describes how well a person can successfully discriminate between two stimuli (Macmillan & Creelman, 2004); in a one-dimensional SDT experiment, d' is equal to the distance between the two perceptual distributions in terms of their root mean square standard deviation, making it a measure of how far apart the two distributions are in terms of how much variance they have. For this experiment, this statistic is extended to work with GRT's two-dimensional representation of perceptual space.

Like SDT, GRT describes each stimulus as producing a distribution of possible perceptual values, and by calculating the distributions' means and variances, a bivariate measure of d' can be calculated. In GRT-wIND, these calculations involve an additional pair of individual difference parameters that scale the perceptual space to best fit each participant's data. These two parameters adjust the variances of all four distributions to either shrink them in both dimensions (larger value for κ lead to less overlap between adjacent distributions, and thus an increased level of performance overall) or shrink the variances in one dimension relative to the other (λ is essentially a ratio for performing better in dimension A than in dimension B). These parameters adjust the scaled perceptual distance between the four distributions without altering their overall structure.

After scaling the GRT-wIND model by these parameters, bivariate d' was defined as the average Mahalanobis distance between a pair of distributions.[4] Mahalanobis distance is a way to measure the distance of a point from

[4]In GRT experiments, it is a recognized problem that there is often no way to equate perceptual distance along the two stimulus dimensions, i.e., it is difficult to say how many units of visual size are equal to one unit of auditory pitch. For simplicity's sake, the units from both stimulus dimensions are being treated as equal for the purposes of calculating d'. While this assumption is unlikely to be perfectly valid, it is also unlikely that small deviations from this assumption would threaten the validity of the analysis, as d' values are only compared to each other within an experiment and both congruent and incongruent d' value are calculated with the same parameters.

the center of a distribution while accounting for the non-zero covariance of that distribution (see De Maesschalck, Jouan-Rimbaud, & Massart, 2000, for a more detailed introduction on Mahalanobis distances). Two estimates of distance are calculated for each pair of stimuli, as each perceptual distribution has its own estimate of covariance, and the participant's bivariate d' value for that pair of stimuli is the mean of those two distances.[5]

The GRT model allows for three types of comparisons. The first type measures sensitivity between compound stimuli that only differ in one dimension. These are labelled 1, 2, 3, and 4 in Figure 24, and are not related to crossmodal correspondence, i.e., the line labelled 1 in the diagram is a measure of how difficult it would be tell apart two stimuli that only differed in their visual component. The second comparison would be between the d' value for the two congruent compound stimuli (line 5) and for the two incongruent stimuli (line 6). If crossmodal correspondences produce an increase in perceptual sensitivity for congruent over incongruent stimuli, that should appear in the data as the d' value for line 5 being larger than line 6. The third type of comparison comes from decomposing the congruent and incongruent d' measures into separate subcomponents along each dimension (these are displayed to the right of the main plot, labelled either **v**isual or **a**uditory). If sensitivity to the visual or auditory components themselves is improved when those components are part of a congruent compound stimulus, it will be reflected in these values.

To evaluate whether perceptual sensitivity was significantly different between congruent and incongruent stimuli, a distribution of congruent-over-incongruent d' values was calculated via bootstrap simulation. While this

[5] An additional note about the individual measures of d': because d' is influenced by the variances of the distributions in question, and the perceptual distributions are the same across participants with the exception of these two parameters, all the variance in different individuals' d' scores will be the result of variance in κ and λ.

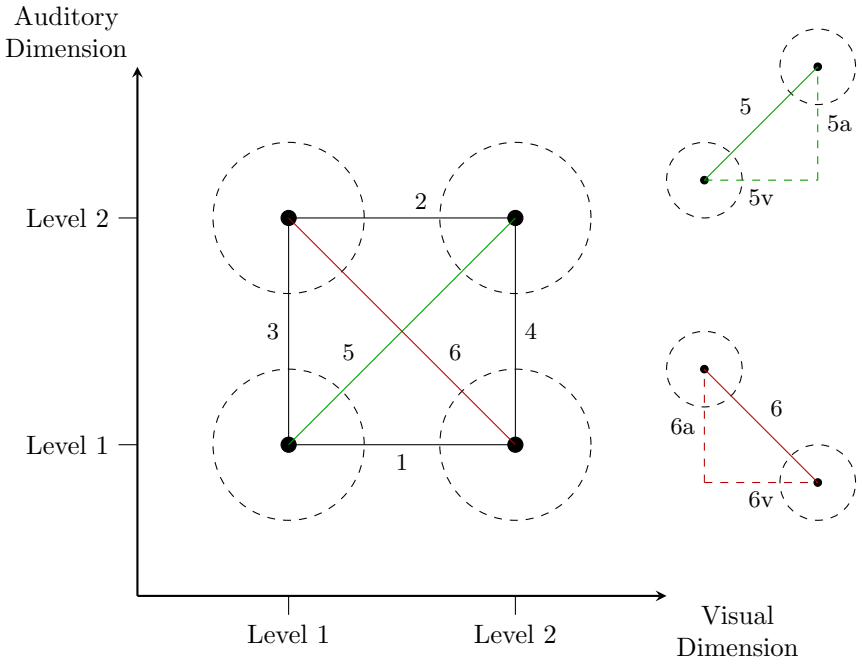

Figure 24. A diagram of the various sensitivity measures that can be calculated in a GRT study, including the visual and auditory components of the crossmodal distances.

strategy is computationally intensive,[6] it allows for confidence intervals around d' estimates to be computed. If the mean bootstrapped d'_{ratio} is statistically significantly larger than 1.0, then it means that there is a significant congruency advantage for crossmodally congruent stimuli.

[6] A single iteration of the simulation can take between 6 and 7 computer core-hours.

References

Alards-Tomalin, D., Leboe-McGowan, J. P., Shaw, J. D., & Leboe-McGowan, L. C. (2014). The effects of numerical magnitude, size, and color saturation on perceived interval duration. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *40*(2), 555.

Algom, D., & Fitousi, D. (2016). Half a century of research on garner interference and the separability–integrality distinction.

Ashby, F. G., & Gott, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(1), 33–53.

Ashby, F. G., & Maddox, W. T. (1990). Integrating information from separable psychological dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 598.

Ashby, F. G., & Soto, F. A. (2014). Multidimensional signal detection theory. In J. R. Busemeyer, J. T. Townsend, Z. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (chap. 2). Oxford University Press.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review*, *93*(2), 154–179.

Belin, P., McAdams, S., Thivard, L., Smith, B., Savel, S., Zilbovicius, M., ... Samson, Y. (2002). The neuroanatomical substrate of sound duration discrimination. *Neuropsychologia*, *40*(12), 1956–1964.

Bernstein, I. H., & Edelstein, B. A. (1971). Effects of some variations in auditory input upon visual choice reaction time. *ournal of Experimental Psychology*, *87*(241).

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433-436.

Chen, Y.-C., & Spence, C. (2010). When hearing the bark helps to identify the dog: Semantically-congruent sounds modulate the identification of masked pictures. *Cognition*, *114*(3), 389–404.

Cohen Kadosh, R., & Henik, A. (2006). A common representation for semantic and physical properties: A cognitive-anatomical approach. *Experimental psychology*, *53*(2), 87-94.

Coward, S. W., & Stevens, C. J. (2004). Extracting meaning from sound: Nomic mappings, everyday listening, and perceiving object size from frequency. *The Psychological Record*, *54*(3), 2.

De Gelder, B., & Bertelson, P. (2003). Multisensory integration, perception and ecological validity. *Trends in cognitive sciences*, *7*(10), 460–467.

Dehaene, S. (1992). Varieties of numerical abilities. *Cognition*, *44*(1), 1–42.

Dehaene, S. (2011). *The number sense: How the mind creates mathematics*. Oxford University Press.

De Maesschalck, R., Jouan-Rimbaud, D., & Massart, D. L. (2000). The mahalanobis distance. *Chemometrics and intelligent laboratory systems*, *50*(1), 1–18.

Evans, K. K., & Treisman, A. (2010). Natural cross-modal mappings between visual and auditory features. *Journal of vision*, *10*(1), 6.

Feigenson, L., Dehaene, S., & Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, *8*(7), 307-314.

Gallace, A., & Spence, C. (2006). Multisensory synesthetic interactions in the speeded classification of visual size. *Perception and Psychophysics*, *68*(7), 1191-1203.

Garner, W. R. (1983). Asymmetric interactions of stimulus dimensions in perceptual information processing. In T. J. Tighe & B. E. Shepp (Eds.), *Perception, cognition, and development: Interactional analyses* (pp. 1–37). Lawrence Erlbaum Hillsdale, NJ.

Garner, W. R., & Morton, J. (1969). Perceptual independence: Definitions, models, and experimental paradigms. *Psychological Bulletin*, *72*(4), 233–259.

Gebuis, T., & van der Smagt, M. J. (2011). Incongruence in number-luminance congruency effects. *Attention, Perception, and Psychophysics*, *73*, 259 – 265.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Hawkins, R. X., Houpt, J., Silbert, N., Blaha, L., & Wickens, T. D. (2016). mdsdt: Functions for analysis of data with general recognition theory [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=mdsdt (R package version 1.2)

Henik, A., & Tzelgov, J. (1982). Is three greater than five: The relation between physical and semantic size in comparison tasks. *Memory and Cognition*, *10*(4), 389-395.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in psychtoolbox-3? *Perception*, *36*(ECVP Abstract Supplement).

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide.* Psychology press.

Marks, L. E. (1987). On cross-modal similarity: Auditory–visual interactions in speeded discrimination. *Journal of Experimental Psychology: Human Perception and Performance*, *13*(3), 384.

Marks, L. E. (1989). On cross-modal similarity: the perceptual structure of pitch, loudness, and brightness. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 586.

Marks, L. E., Ben-Artzi, E., & Lakatos, S. (2003). Cross-modal interactions in auditory and visual discrimination. *International Journal of Psychophysiology*, *50*(1), 125–145.

MATLAB. (2012). *Matlab and statistics toolbox release 2012b* (2012b ed.). Natick, Massachusetts, United States.: The MathWorks, Inc.

Oliveri, M., Vicario, C. M., Salerno, S., Koch, G., Turriziani, P., Mangano, R., ... Caltagirone, C. (2008). Perceiving numbers alters time perception. *Neuroscience letters*, *438*(3), 308–311.

Parkinson, C., Kohler, P. J., Sievers, B., & Wheatley, T. (2012). Associations between auditory pitch and visual elevation do not depend on language: Evidence from a remote population. *Perception*, *41*(7), 854–861.

Patching, G. R., & Quinlan, P. T. (2002). Garner and congruence effects in the speeded classification of bimodal signals. *Journal of Experimental Psychology: Human Perception and Performance*, *28*(4), 755.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437-442.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from http://www.R-project.org/

Sanabria, D., Spence, C., & Soto-Faraco, S. (2007). Perceptual and decisional contributions to audiovisual interactions in the perception of apparent motion: A signal detection study. *Cognition*, *102*(2), 299–310.

Scheid, S., & Kubovy, M. (in preparation). A new method for cross–modal research: The incompleteness of our current paradigms.
(Unpublished Manuscript)

Soto, F. A., Vucovich, L., Musgrave, R., & Ashby, F. G. (2015). General recognition theory with individual differences: a new method for examining perceptual and decisional interactions with an application to face perception. *Psychonomic bulletin and review*, *22*(1), 88-111.

Soto, F. A., & Zheng, E. (2016). grtools: General recognition theory tools for the analysis of perceptual independence [Computer software manual].

Soto, F. A., Zheng, E., Fonseca, J., & Ashby, F. G. (2016). Testing separability and independence of perceptual dimensions with general recognition theory: A tutorial and new r package (grtools) [Computer software manual].

Spence, C. (2011). Crossmodal correspondences: A tutorial review. *Attention, Perception, and Psychophysics*, *73*(4), 971-995.

Spence, C., & Parise, C. V. (2012). The cognitive neuroscience of crossmodal correspondences. *i-Perception*, *3*(7), 410.

Stevens, S. S. (1957). On the psychophysical law. *Psychological review*, *64*(3), 153.

Szűcs, D., & Soltész, F. (2008). The interaction of task-relevant and task-irrelevant stimulus features in the number/size congruency paradigm: An erp study. *Brain Research*, *1190*, 143–158.

Wagner, S., Winner, E., Cicchetti, D., & Gardner, H. (1981). " metaphorical" mapping in human infants. *Child Development*, 728–731.

Walker, A. R. (1985). Mental imagery and musical concepts: Some evidence from the congenitally blind. *Bulletin of the Council for Research in Music Education*, 229–237.

Walsh, V. (2003). A theory of magnitude: common cortical metrics of time, space and quantity. *Trends in cognitive sciences*, *7*(11), 483–488.